A Basic Language Technology Toolkit for Quechua

A thesis submitted in fulfilment of the requirements
for the degree of Doctor of Philosophy

to the

Faculty of Arts of the University of Zurich

by

Annette Rios

Accepted in the Autumn Term 2015
on the Recommendation of the Doctoral Committee:

Prof. Dr. Martin Volk (main advisor)
Prof. Dr. Balthasar Bickel
Zurich, 2015

Abstract

In this thesis, we describe the development of several natural language processing tools and resources for the Andean language Cuzco Quechua as part of the SQUOIA project at the University of Zurich.

The main focus of this work lies on the implementation of a machine translation system for the language pair Spanish-Cuzco Quechua. Since the target language Quechua is not only a non-mainstream language in the field of computational linguistics, but also typologically quite different from the source language Spanish, several rather unusual problems became evident, and we had to find solutions in order to deal with them. Therefore, the first part of this thesis presents monolingual tools and resources that are not directly related to machine translation, but are nevertheless indispensable.

The main contributions of this thesis are as follows:

- We built a hybrid machine translation system that can translate Spanish text into Cuzco Quechua. The core system is a classical rule-based transfer engine, however, several statistical modules are included for tasks that cannot be resolved reliably with rules.
- We implemented a text normalization pipeline that automatically rewrites Quechua texts in different orthographies or dialects to the official Peruvian standard orthography. This includes a tool for the morphological analysis of Quechua words that achieves high coverage. Furthermore, we also created a slightly adapted version that can be used as spell checker back-end, in combination with a plug-in for the open-source productivity suite LibreOffice/OpenOffice.
- We built a Quechua dependency treebank of about 2000 annotated sentences, that provided not only training data for some of the translation modules, but also served as a source of verification, since it allows to observe the distribution of certain syntactic and morphological structures. Furthermore, we trained a statistical parser on the treebank and thus have now a complete pipeline to morphologically analyze, disambiguate and then parse Quechua texts.

All resources and tools are freely available from the projects website.[1]

Apart from the scientific interest in developing tools and applications for a language that is typologically distant from the main stream languages in computational linguistics, we hope that the various resources presented in this thesis will be useful not only for language learners and linguists, but also to Quechua speakers who want to use modern technology in their native language.

[1] https://github.com/ariosquoia/squoia

Acknowledgements

Above all, I would like to thank my supervisor Martin Volk for his support and guidance during the four years of this project. I am also very grateful for the continued assistance, endless discussions and many laughs with my fellow researcher in the SQUOIA project, Anne Göhring. I would also like to thank the members of the doctoral committee, Balthasar Bickel and Paul Heggarty, who provided a detailed review with many suggestions for improvement.

Moreover, I wish to thank the people in Peru that made this work possible:

- Richard Castro Mamani for the collaboration on the spell checkers, the management and organization of the evaluation of the MT system and the translations for the treebank
- Roger Gonzalo Segura for the syntactic annotation and the numerous discussions about Quechua syntax
- César Morante Luna for translations, corrections and filling the gaps of the bilingual dictionary of the MT system
- Virginia Mamani Mamani and Irma Álvarez Ccoscco for the contribution of the translations of the treebank texts
- Juan Cruz Tello for providing contacts and general support of the project

Furthermore, I would like to thank all my colleagues at the Institute of Computational Linguistics, especially Simon Clematide for the provided help with the finite-state tools and my fellow PhD students Magdalena Plamada, for general advice on MT related issues, Don Tuggener for ideas and discussions about coreference resolution to deal with Quechua switch-reference, Johannes Graën for the technical support with the web-related parts of this thesis, and my former colleague Rico Sennrich for his valuable tips and tricks concerning the machine learning parts of the Spanish-Quechua translation system.

I would also like to thank my family, especially Naira and my mother Susanne for their patience and support during these past four years.

Most importantly, I am grateful for the financial support provided by the Swiss National Science Foundation under grants 100015_132219 and 100015_149841.

Contents

List of Figures

List of Tables

Abbreviations

Glosses (morphology)		Quechua suffixes
ABL	ablative	*-manta*
ACC	accusative	*-ta*
ADD	additive	*-pas/-pis*
AFF	affective	*-yku*
AG	agentive	*-q*
ASMP	asumptive	*-cha/ch*
AUTOTRS	autotransitive	*-ya*
BEN	benefactive	*-paq*
CAS	case suffix	
CAUS	causative	*-chi*
COMP	complementizer (Spanish)	
CON	connective	*-taq*
CONT	continuative	
DAT	dative	*-man*
DES	desiderative	*-naya*
DEF	definitive	*-puni*
DIM	diminutive	*-cha*
DIR	directional	*-mu*
DIRE	direct evidential	*-mi/-m*
DISTR	distributive	*-kama/-nka*
DS	different subject	*-pti*
EXCL	exclusive (in first plural person)	
FACT	factitive	*-cha*
FOC	focus	

FUT	future tense	
GEN	genitive	*-pa/-p*
HAB	habitual past	
IMP	imperative	
INCH	inchoative	*-ri*
INCL	inclusive (in first plural person)	
INDE	indirect evidential	*-si/-s*
INF	infinitive	*-y*
INSTR	instrumental	*-wan*
INTR	interrogative	*-taq/-chu*
IPST	past tense, indirect evidentiality	*-sqa*
LIM	limitative	*-lla*
LOC	locative	*-pi*
NEG	negation	*-chu*
NPers	nominal person suffix	
NRoot	nominal root	
NS	nominalizing suffix	
OBL	obligative	*-na*
OBJ	object	
PERF	perfect	*-sqa*
PREP	preposition (Spanish)	
PRES	present tense	
PL	plural	*-kuna*
POSS	possessive	
PROG	progressive	*-chka*
PST	Past	*-rqa*
REL	relative pronoun (Spanish)	
RFLX	reflexive	*-ku*
RPTN	repentine	*-rqu*
SG	singular	
SS	same subject	*-spa*
TERM	terminative	*-kama*
TOP	topic	*-qa*

VDeriv	verbal derivational suffix	
VDIM	verbal diminutive	*-cha*
VPers	verbal person (subject or object)	
VRoot	verbal root	
VS	verbalizing suffix	

Edge labels (dependency treebank)

adv	adverbial modifier
co	coordination
det	determiner/demonstrative
ev	evidential
hab	habitual past
iobj	indirect object
loc	location
mod	unspecified modfier
neg	negation
ns	nominalization
obj	direct object
poss.subj	subject (possessor, genitive)
pred	predicative element
punc	punctuation
qnt	quantifier
rep	repeated element
s.arg	argument of suffix
s.co	coordinative suffix
s.neg	negation suffix
s.poss.subj	subject (possessive suffix)
s.subj	subject (suffix)
s.subj_iobj	subject and indirect object suffix
s.subj_obj	subject and direct object suffix
subj	subject
tmp	temporal modifier
VROOT	virtual root

Chapter 1

Introduction

1.1 Overview

Even though the availability of texts in digital form has significantly increased over the last few decades, a wide gap between rich and poor nations in terms of access to this information still persists [Kshetri and Nikhiles 2009]. An important factor for this *Global Digital Divide* is a lack of infrastructure in poor countries, but there is also a linguistic dimension to the problem: language plays a fundamental role in this development, since people interact with the new technologies mainly using language [Gasser 2006:1-2]. Although the exact number of websites per language is unknown, it is beyond dispute that the vast majority of the digital information in the world wide web is accessible in only a small number of languages, most notably English. Furthermore, technological solutions suffer from a whole range of cultural and linguistic biases, as software interfaces, programming languages and even markup languages are heavily influenced by English. For most of the world's languages, even the most basic language technology applications, such as spell checkers or machine readable dictionaries, have not been developed, due to the fact that focusing on software for the languages of the Global North is clearly more profitable [Gasser 2006].

All these circumstances lead to what Gasser [2006] calls the *Linguistic Digital Divide*, defined by the following aspects:

- *knowledge gap*: the lack of digital information written in disadvantaged languages

- *participation gap*: the relative lack of input from the disadvantaged linguistic communities in the global decision making
- *software gap*: the lack of computational tools that facilitate the integration of speakers of disadvantaged languages into the digital world

The focus of this thesis lies on how to bridge the software gap for Cuzco Quechua, a language that has very few digital resources, even though it is one of the largest indigenous languages in the Americas by number of speakers.

Some of the tools and applications presented in the following chapters were planned from the beginning, while others were created in order to deal with special circumstances. For instance, the need for a statistical language model trained on Quechua texts lead to the implementation of a normalization pipeline, since the performance of statistical models is severly impaired with non-standardized text written in different orthographies and/or dialectal varieties.

Furthermore, the development of tools for a non-mainstream language revealed gaps in common NLP approaches: even though the rather exotic features of Quechua, such as evidentiality or internally headed relative clauses, are well known in the linguistic community and have been described in detail, they are rarely dealt with in computational linguistics, since they are absent in the commonly treated languages. Apart from these intriguing technical issues, further motivation for this work comes from the hope that the developed tools and resources will help language learners and native speakers alike in processing Cuzco Quechua text, and might in this sense help to preserve and support the language itself.

1.2 The Quechua Language Family

Quechua is a group of closely related languages, spoken by 8-10 million people in Peru, Bolivia, Ecuador, Southern Colombia and the North-Western parts of Argentina. Ethnologue[1] also lists a small number of Quechua speakers for Chile. Quechua is one of the official languages of Peru, Bolivia and Ecuador.

[1] http://www.ethnologue.com

Although Quechua is often referred to as a 'language' and its local varieties as 'dialects', Quechua represents a language family, comparable in depth to the Romance or Slavic languages [Adelaar and Muysken 2004:168]. Mutual intelligibility, especially between speakers of distant dialects, is not always given.

1.2.1 Distribution of Quechua Languages

Traditionally, the Quechua languages have been divided into two main branches, Quechua I and II in terms of the Peruvian linguist Torero [1964], respectively Quechua A and B in terms of the American linguist Parker [1963], see Fig. 1.1 for the geographic distribution. Even though more recent studies [Heggarty 2005, Landerman 1991] suggest that this binary classification might not be accurate, we will use Torero's labels in this introduction for reasons of simplicity. However, this choice does by no means imply a particular preference for Torero's classification, in fact, the topic of Quechua dialect classification goes far beyond the scope of this introduction.

Quechua I is the more archaic group of dialects, spoken in Central Peru. It comprises a heavily fragmented dialect complex with limited mutual comprehension between the different local varieties, although they share a number of common features [Adelaar and Muysken 2004:185]. This area is probably the Quechua homeland, the place from where the language family originally spread out [Cerrón-Palomino 2003].

The second branch, Quechua II, comprises all the remaining Quechua dialects:

- QIIA, spoken in Northern Peru
- QIIB, spoken in Ecuador and Colombia
- QIIC, spoken in Southern Peru, Bolivia, and Argentina

The dialects of Quechua IIA occupy an intermediate position between Quechua I and the rest of Quechua II [Adelaar and Muysken 2004:186]. The classification of these dialects is not as straightforward as it might seem: the northern dialects of Cajamarca and Ferreñafe have attributes of both Quechua IIB and Quechua I, whereas the dialects of Yauyos hold a similar position between the Quechua IIC and Quechua I varieties [Adelaar and Muysken 2004:186], and in fact, the status of this dialect group is one of

the most debated issues concerning the classification of Quechua varieties.

Quechua IIB, comprises the Ecuadorian branch (*Kichwa*), the Quechua spoken in Colombia (*Inga* or *Ingano*) and the dialects spoken in the Peruvian departments of San Martín, Loreto and Amazonas [Adelaar and Muysken 2004:187].

Quechua IIC comprises all the remaining Quechua dialects to the south of the Quechua I group, including the dialect groups Ayacucho, Cuzco-Bolivia and Argentina. The division between Ayacucho and Cuzco-Bolivian Quechua is mainly due to the occurrence of glottalized and aspirated stops in the Cuzco-Bolivian dialects, a phonetic distinction that Ayacucho and Argentina Quechua lack. However, Cuzco-Bolivian Quechua itself is no homogeneous group at all [Adelaar and Muysken 2004:187],[Cerrón-Palomino 2003:242-245].

1.3 NLP for Quechua

Although Quechua is clearly a low-resource language, it has to be noted that several groups have been working on computational linguistic applications for Quechua over the past few years.

The *Instituto de Lengua y Literatura Andina Amazónica* (ILLA)[3] in La Paz, Bolivia has been working for several years on the digitization of dictionaries, books and grammars. Furthermore, they translate user interfaces of open-source software into the native languages of Bolivia, such as Quechua and Aymara. Several electronic dictionaries for GoldenDict and StarDict are freely available from their website. Additionally ILLA offers SimiDic, an electronic dictionary for Android phones that includes vocabulary in Quechua, Aymara and Guaraní.

[2]Map from Open Source Maps at: http://www.mapsopensource.com/south-america-countries-outline-map-black-and-white.html, dialect distribution according to Adelaar and Muysken [2004]

[3]http://www.illa-a.org/wp/illa/

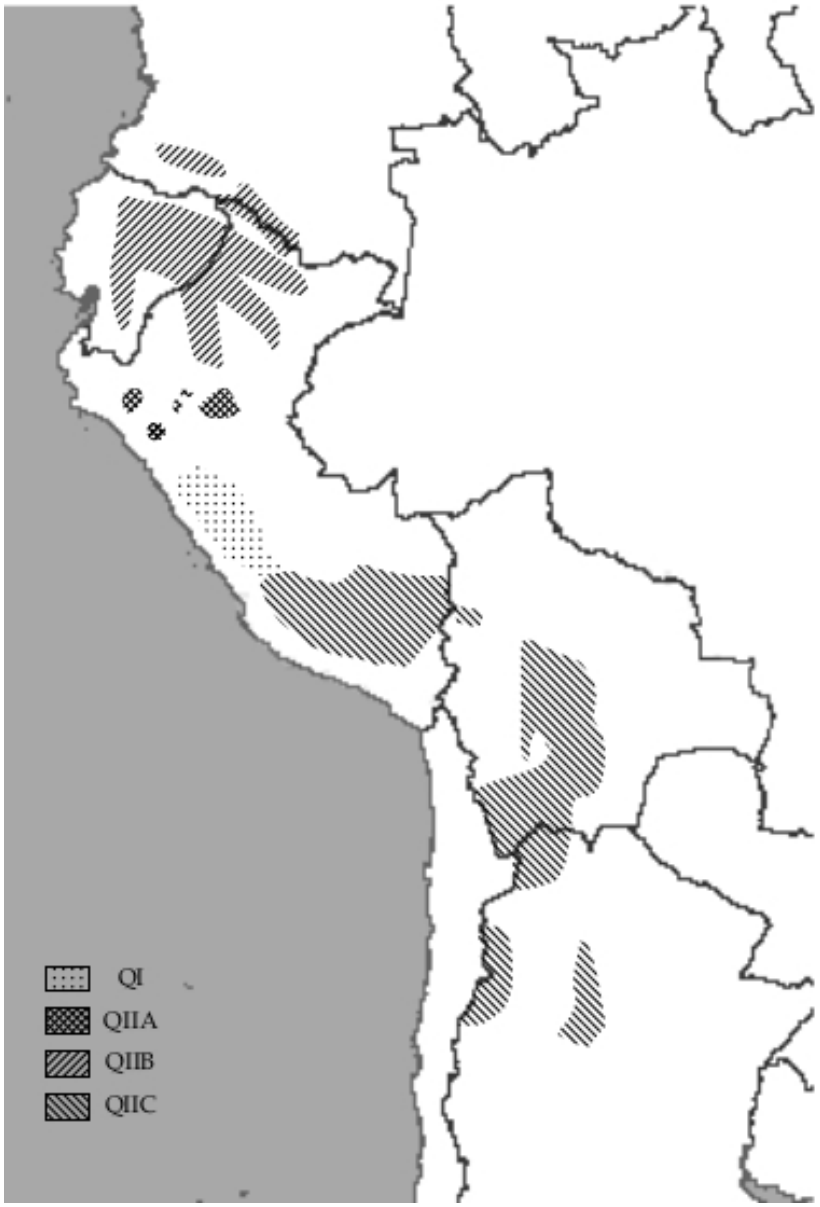


FIGURE 1.1: Quechua Dialect Groups[2]

In Peru, the most notable contribution to NLP for indigenous languages and especially Quechua comes from the group *hinantin* at the *Universidad Nacional San Antonio Abad del Cusco* (UNSAAC). Their work so far includes a text-to-speech system for Cuzco Quechua, an interactive grammar for language learners and graphical web-interfaces for the available Quechua spell checkers. Furthermore, one of the PhD students involved in the project, Hugo Quispe, is working on the creation of a lexical database for Cuzco Quechua.[4]

[4] http://hinantin.com/home2/proyectos/base-datos-lexica.html

As for machine translation, there is a surprisingly large number of MT projects that include Quechua. The open-source MT platform Apertium has a Quechua-Spanish translation system in *nursery* (early stage of development), originally implemented by Czech linguist Vlastimil Rataj. Following his work, a group centered at the *Universidad Nacional Micaela Bastidas* in Abancay, Peru, has adapted the original Quechua-Spanish system to the translation direction Spanish-Quechua [Larico Uchamaco et al. 2013].

Quechua is also one of the languages involved in the project *Human Language Technology and the Democratization of Information*, short *hltdi-l3*[5], that aims at developing MT systems for under-resourced languages of the Global South. Apart from a freely available collection of corpora, they also offer a morphological analyzer, antiMorfo, in combination with a user interface for morphology learning [Gasser 2011].

AVENUE was another MT project that included Quechua. The main idea in AVENUE was to create parallel corpora for low-resource languages through elicitation [Monson et al. 2006]. AVENUE is not actively maintained anymore, but part of their resources are available from the *hltdi-l3* project's website.[6]

Furthermore, Mohler and Mihalcea [2008] made an experiment with statistical machine translation Spanish-Quechua in order to test Babylon, a tool that gathers parallel text from the internet. They achieved an improvement of several BLEU points with the crawled data as opposed to the baseline system trained on parts of the Bible. The best result for Spanish to Quechua was 6.42 BLEU points, whereas the highest score for Quechua to Spanish was 8.02 BLEU points. It has to be noted though that they did not distinguish between the different Quechua languages, they used texts in all varieties and orthographies for their system.

1.4 The SQUOIA Project

All applications and resources described in this thesis have been developed in the SQUOIA project[7] at the University of Zurich, funded by the Swiss National Science

[5] `http://www.cs.indiana.edu/~gasser/Research/projects.html`

[6] `https://code.google.com/p/hltdi-l3/wiki/AvenueQuechuaCorpus`

[7] SQUOIA: ***S****panish to* ***Qu****echua translation through exploi****t****ation of parallel treebanks*, see also `http://www.cl.uzh.ch/research/maschinelleuebersetzung/hybridmt_en.html`.

Foundation under grants 100015_132219 and 100015_149841. The main focus of the project was to investigate how parallel treebanks can be exploited for machine translation, and how the availability of resources for different language pairs impacts the development of MT systems. For this reason, we built hybrid machine translation systems for Spanish-Quechua and Spanish-German, the former as a case study for a system with a resource-scarce target language and the latter as an instance of a resource-rich language pair. Since the source language for both translation systems is Spanish, we could in part use the same tools for both language pairs.

The main part of my work over the past 4 years was the development of the translation system Spanish-Cuzco Quechua. However, the special situation of the target language Quechua as a non-mainstream language in computational linguistics and as a low-prestige language in society resulted in many language specific problems that had to be solved along the way. For instance, the wide range of different orthographies used in written Quechua[8] is a problem for any statistical text processing, such as the training of a language model to rank different translation options. Thus, in order to get a statistical language model, we had to first find a way to normalize Quechua texts automatically. Furthermore, the resulting normalization pipeline can be adapted for spell checking with little effort. For this reason, an entire chapter of this thesis is dedicated to the treatment of Quechua morphology: although not directly related to machine translation, automatic processing of Quechua morphology provides the necessary resources for important parts of the translation system.

Furthermore, the project involved the creation of a parallel treebank in all 3 languages. While the Spanish and German parts of the treebank were finished within the first year of the project, building the Quechua treebank took considerably longer: before the actual annotation process started, the texts had to be translated into Quechua. Additionally, we had to design an annotation scheme from scratch and set up pre-processing and annotation tools. A by-product of the resulting parallel corpus is the Spanish-Quechua version of Bilingwis, a web tool that allows to search for translations in context in word-aligned texts.

[8] See Table 2.3 on page 17 in the following chapter for examples of different Quechua orthographies.

Generally speaking, the monolingual tools and resources described in the first part of this thesis are necessary to build the multilingual applications of the second part.

1.5 Research Questions

This thesis will focus on the following questions regarding the development of NLP tools for a language with scarce resources and issues related to this undertaking:

1. How much do we have to rely on rules in NLP applications that involve a low-resource language and can we still make use of statistics?
2. What implications does a complex agglutinative morphology have for the development of NLP applications and to what extent do we need to adapt common approaches?
3. What issues arise in a machine translation system with typologically distinct source and target language that encode different grammatical categories?
4. What are the essential resources to build a machine translation system with a low-resource target language and how can they be created efficiently?

1.6 Thesis Outline

In this chapter, we have set the broader context for the development of NLP tools for a low-resource language and we gave a short overview on characteristics and the distribution of the Quechua languages. The rest of this thesis is structured as follows:

Chapter 2

Morphology

This chapter explains the morphological structures in Quechua word formation and how we deal with them in technological applications.

Chapter 3

Treebanks

This chapter summarizes the treebanking guidelines and how existing tools were adapted

for the syntactic annotation of Quechua texts.

Chapter 4

Bilingwis

This chapter describes how Bilingwis, an online service for searching translations in parallel, word-aligned texts, was adapted to the language pair Spanish-Quechua.

Chapter 5

Machine Translation

This chapter describes in detail the implementation of a machine translation system for the language pair Spanish-Quechua, with a special focus on resolving morphological ambiguities.

Part I

Monolingual Quechua Resources

Chapter 2

Quechua Morphology

2.1 Introduction

This chapter presents an in-depth introduction to Quechua morphology and the automatic analysis and disambiguation of Quechua words. For readers not interested in the technical details, sections 2.3 and 2.4 are not essential to understand the following chapters.

All Southern Quechua dialects are strongly agglutinative, suffixating languages.[1] There are over 130 suffixes, although the exact number and morphotactical behavior may differ to some extent even within the Southern Quechua dialect group. Varieties spoken close to or overlapping with Aymara speaking regions have borrowed not only roots, but also a small number of suffixes. For instance, the Quechua spoken in Puno includes Aymara suffixes such as e.g. *-thapi*, *-t'a* or *-naqa*[2] that are unknown in other dialects [Adelaar 1987]. Apart from lexical differences, suffixes can also have different forms across dialects, e.g. the progressive in Ayacucho is marked by *-chka*, in Cuzco by *-sha*, and in Bolivia by *-sa* or *-sya*. Furthermore, the order of suffixes in combinations can vary to a certain degree, even within one text.

[1] Parts of this chapter are based on Rios[2011a], Rios[2011b] and Rios and Castro Mamani [2014].

[2] *-thapi*: directional that indicates movement towards each other, concentration, contraction; *-t'a*: denotes a unique, short or instantaneous action; *-naqa*: diffuse, aimless action, often with movement verbs [Adelaar 1987]

plain	glottalized	aspirated
ch	ch’	chh
k	k’	kh
p	p’	ph
q	q’	qh
t	t’	th
tanta ‘together’	*t’anta* ‘bread’	*thanta* ‘old’ (things)

TABLE 2.1: Glottalization and Aspiration

Apart from lexical differences, there is one major dialectal distinction between the Cuzco and Bolivian dialects on one side, and the Ayacucho and Argentina varieties on the other side: Cuzco and Bolivian Quechua have, like Aymara, a three way distinction of stops (plain, glottalized and aspirated), whereas Ayacucho and Argentina Quechua have only plain stops, see Table 2.1.

Additionally, there are some morphotactic differences concerning the combination of suffixes: for instance, a small number of Quechua suffixes change their vowel from *u* to *a* in combination with certain suffixes, but the exact contexts that induce this vowel change differ slightly across dialects.

There are five functional classes of Quechua suffixes (see Table 2.2): nominalizing (noun→verb) and verbalizing (verb→noun), nominal (noun→noun) and verbal (verb→verb) suffixes and so-called independent or ambiguous suffixes, that can be attached to both verbal or nominal forms, without altering the part of speech. The position of these suffixes is at the end of the suffix sequence, their relative order is more or less fixed, though dialects show minor variations. The functions of the independent suffixes include data source, polar question marking and topic or contrast, amongst others [Adelaar and Muysken 2004:208].

Figure 2.1 contains a simplified illustration of Quechua word formation: a word can start with a nominal root, followed by nominal suffixes and optional independent suffixes, or it may start with a verbal root, followed by verbal suffixes and optional independent suffixes. Furthermore, it is possible and quite frequent to switch between these two paradigms: a verb can be nominalized even if suffixes from the verbal slots 1-3 (derivation, object, aspect) are present, see example (1). Nouns, on the other hand, cannot be

1	nominalizing	V → N
	qillqa -q ,'write-AG' ⇒	writer
2	verbalizing	N → V
	llamp'u -ya-, 'soft-AUTOTRS' ⇒	to become soft
3	nominal (derivation/inflection)	N → N
	wasi -yuq , 'house-POSS' ⇒	house owner
4	verbal (derivation/inflection)	V → V
	wañu -chi- , 'die-CAUS' ⇒	kill
5	independent	*e.g. evidentiality, topic, epistemic modality...*

TABLE 2.2: Suffix Classes

verbalized if nominal suffixes are attached. This process of verbalization and nominalization can be repeated within a word, see example (2).[3]

(1) *haywa -yku -wa -na -nku -paq*
to.hand -AFF -1.OBJ -OBL -3.PL.POSS -BEN
VRoot VDeriv VPers NS NPers Case
'in order for them to hand me [something]'

(2) *yupa -y -cha -y*
count -INF -FACT -INF
VRoot NS VS NS
'respect'
(lit. 'the-to-make-count')

A Quechua word may have more than one case suffix, although possible combinations are restricted to a relatively small number and may have a non-compositional meaning, see examples (3) and (4).

(3) *karu* ***-pi*** ***-kama*** *-m*
far -LOC -DISTR -DIRE
yacha -nku.
live -3.PL
'They live far away from each other.'

(4) *tayta -y* ***-pa*** ***-ta*** *ri*
father -1.SG.POSS -GEN -ACC go
-chka -ni.
-PROG -1.SG
'I'm going to my father's [place].'

[Soto Ruiz 1976:82-83]

[3] All Quechua examples in this thesis will be presented in the standard orthography [Cerrón-Palomino 1994] for simplicity. For a list of the abbreviations used in the glosses, see xiii.

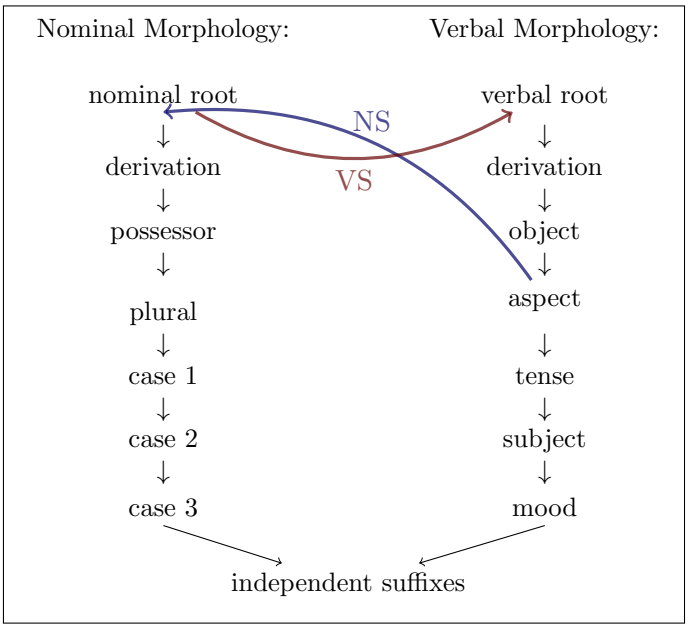


FIGURE 2.1: Quechua Word Formation (simplified)

Quechua roots are, apart from a small number of particles, either verbal or nominal. Adjectives do not constitute a word class on their own on a morphological level, as they behave exactly the same as nominal roots. There may be some syntactic restrictions on true adjectives [Adelaar and Muysken 2004:208], but these can be ignored for a automatic morphological analysis. Many roots are indeed ambiguous and can be used either as noun or verb without any derivational suffixes:

(5) *taki* *-y*
song/sing -1.SG.POSS
'My song'

(6) *taki* *-ni.*
song/sing -1.SG
'I sing'

2.2 Orthographic Variation

Apart from the dialectal differences, there is also a wide range of orthographic variation within the Southern Quechua dialect group. Several standards have been proposed, most notably the standardized orthography as defined by Cerrón-Palomino [1994]. This standard has been adopted by the Bolivian government [Quiroz Villarroel 2000], with one small adaption: in Bolivia, the glottal fricative [h] is written as ⟨j⟩ instead of ⟨h⟩. In Peru, the situation is slightly more complicated: although the Ministry of Education has

AMLQ	*mana qelqaq yachaq ñausa qelqa runasimipi kasqanku rayku [..]*
norm.	mana qillqaq yachaq ñawsa qillqa runasimipi kasqankurayku [..]
span.	*Cay teccsimuyuta, hanacc-pachatapas, Ccanmi tacyachinqui, Ccanmi ticrachinqui [..]*
norm.	Kay tiqsimuyuta, hanaq pachatapas, Qammi takyachinki, Qammi t'ikrachinki [..]
boliv.	*Chaywampis paykuna onqosqa kashajtinku, noqaqa llakiy qhashqa p'achasta churakorqani.*
norm.	Chaywanpas paykuna unqusqa kachkaptinku, ñuqaqa llakiy qhachqa p'achakunata churakurqani.

TABLE 2.3: Different Orthographies with Corresponding Standardized Version[5]

defined an official standard orthography[4],there is still some disagreement regarding the correct spelling of Quechua words. The main opponent to the Ministry of Education is the Academia Mayor de la Lengua Quechua (AMLQ) in Cuzco that uses and propagates their own orthography. There are two main problems with their standard: first of all, it is based exclusively on the regional dialect of Cuzco, ignoring any dialectal characteristics of other varieties. Secondly, their orthography is in part phonetic, e.g. due to fricativization of syllable final plosives, they write *akllay* as *ajllay*, and *raphra* as *rafra*. Furthermore, they use 5 vowels (a,e,i,o,u) although only three of them are phonemic, as [e] and [o] are allophones of /i/ and /u/ that occur only in proximity to postvelar /q/.

In addition to the different writing standards, many Quechua texts are written in a more or less Spanish orthography, where for instance [wa] is written as ⟨hua⟩, and [ki] is written as ⟨qui⟩. Table 2.3 illustrates the orthography of the Academia Mayor de la Lengua Quechua in Cuzco (first row), a typical 'Spanish' spelling (second row) and an old, non-standardized Bolivian spelling (last row), as opposed to the unified standard orthography defined by Cerrón-Palomino [1994].

2.3 Morphological Analysis

The first step in the automatic normalization is to split the Quechua word forms into morphemes, and since certain morphemes or morpheme sequences are ambiguous, the

[4]As declared in the *Resolución Ministerial N° 1218-85-ED de 1985*

[5]AMLQ = Academia Mayor de la Lengua Quechua en Cusco, norm = normalized, span = Spanish orthography, boliv = (old) Bolivian orthography

second step is to find the correct analysis if more than one is possible. Section 2.3.1 will introduce an efficient and fast method for the morphological segmentation of words, the so-called finite-state networks. In section 2.3.2, we will present the implementation of the finite-state analyzer for Quechua that serves as basis for the automatic normalization.

2.3.1 Finite-State Networks

In formal language theory, a language is simply a set of words, and as such can undergo set operations like union, intersection or substraction. The words of a formal language are composed of a finite set of symbols, i.e. letters, of an alphabet Σ. For the treatment of morphology in most natural languages, the simplest and most restrictive type of formal languages, the so-called regular languages, are sufficient.

Formally, a regular language over an alphabet Σ is defined as follows [Carstensen et al. 2010:71]:

- The empty language $\emptyset$ and the language that contains only the empty word $\{\epsilon\}$ are regular languages
- For each $a \in \Sigma$, the singleton $\{a\}$ is a regular language
- If L1 and L2 are regular languages, then
  - union: L1 $\cup$ L2
    if L1=$\{a\}$ and L2=$\{b\}$, then L1 $\cup$ L2 = $\{a, b\}$
  - concatenation: L1 $\cdot$ L2
    if L1=$\{a\}$ and L2=$\{b\}$, then L1 $\cdot$L2 = $\{ab\}$
  - Kleene star: L1*
    if L1=$\{a\}$, then L1*= $\{a, aa, aaa, aaaa..\}$

  are also regular languages
- no other languages over Σ are regular languages

Such a regular language, which contains all grammatical words consisting of symbols from Σ, can be represented as a finite-state network: a graph that consists of nodes (states) linked together with directed arc-transitions. The transitions are labeled with symbols from Σ, in a manner such that every word of the language corresponds to a

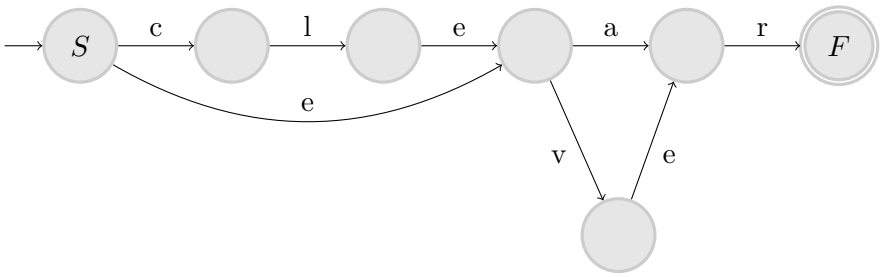


FIGURE 2.2: Finite-State Machine with L= {*clear,clever,ear,ever*} [Beesley and Karttunen 2003:17]

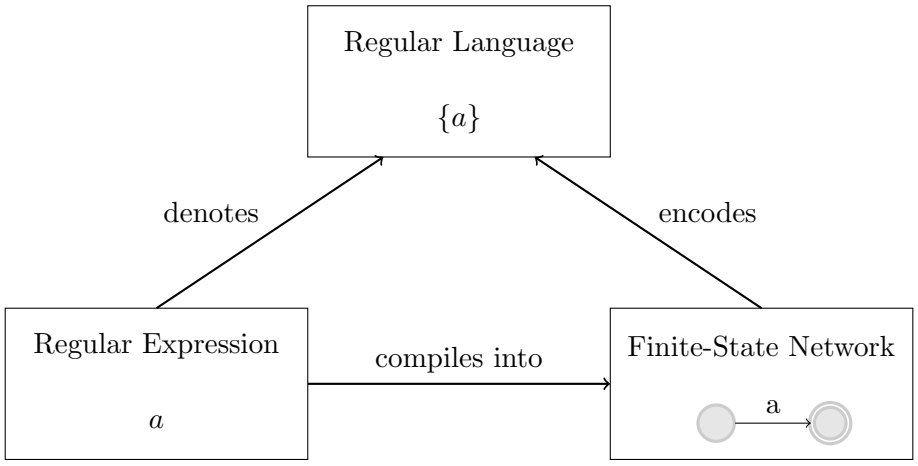


FIGURE 2.3: Interdependence Regular Expression - Language - Network [Beesley and Karttunen 2003:44]

path in the network [Beesley and Karttunen 2003:43]. See Fig. 2.2 with the finite-state network for the language L={*clear*, *clever*, *ear*, *ever*}, taken from [Beesley and Karttunen 2003:17]. Final states are by convention drawn as double-lined circles.

A formal notation to describe a regular language are regular expressions, which in turn can be compiled into a finite-state machine. Since the resulting finite-state machine contains all the words from the given regular language and no other words, the finite-state machine can decide if a given word is part of the regular language it encodes. Figure 2.3 illustrates the interdependence between regular languages, regular expressions and finite-state networks [Beesley and Karttunen 2003:44].

Furthermore, there is a distinction between deterministic and non-deterministic finite-state machines: in a deterministic finite-state machine, every state has exactly one transition for each possible input symbol, while in a non-deterministic finite-state machine the input of a particular symbol can lead to more than one transition for a given

state.

A finite-state machine M=$\langle \Phi, \Sigma, \delta, S, F \rangle$ consists of [Carstensen et al. 2010:74-75]:

- a set of states: Φ
- a start state: $S \in \Phi$
- a set of final states: $F \subseteq \Phi$
- a set of symbols (alphabet): Σ
- state transition function:
  deterministic FSM: $\delta : S \times \Sigma \rightarrow \Phi$
  non-deterministic FSM: $\delta : S \times \Sigma \rightarrow \wp(\Phi)$[6]

There is an important distinction between finite-state machines (FSM) that are one-sided, and finite-state transducers (FST) that have an upper and a lower side, or more generally, an input and an output level. Since a finite-state transducer has two sides, it can not only decide if a given word is part of its regular language, but it will also return the corresponding output to the given input [Beesley and Karttunen 2003:11].

A finite-state transducer accordingly implements a relation between two regular languages: an upper-side and a lower-side regular language, and it literally 'transduces' strings from one language into the other. In a non-deterministic finite-state transducer, it may produce more than one possible output for a given string.

A finite-state transducer T=$\langle \Phi, \Sigma, \Gamma, \delta, S, F \rangle$ consists of [Carstensen et al. 2010:78-79]:

- a set of states: Φ
- a start state: $S \in \Phi$
- a set of final states: $F \subseteq \Phi$
- a set of input symbols: Σ
- a set of output symbols: Γ
  deterministic FST: $\delta : S \times (\Sigma \cup \{\epsilon\}) \times (\Gamma \cup \{\epsilon\}) \rightarrow \Phi$
  non-deterministic FST: $\delta : S \times (\Sigma \cup \{\epsilon\}) \times (\Gamma \cup \{\epsilon\}) \rightarrow \wp(\Phi)$

[6] $\wp(\Phi)$ = power set of Φ, i.e. the set of all subsets of Φ.

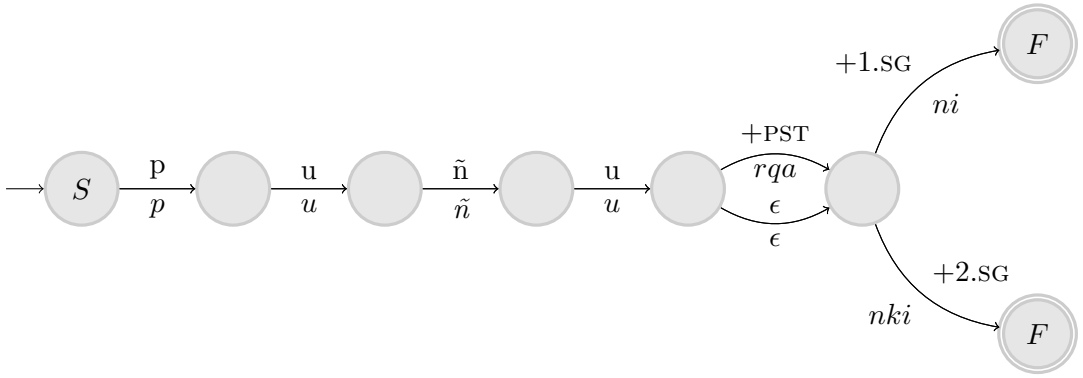


FIGURE 2.4: Finite-State Transducer with *puñu-*

See Fig. 2.4 for an example of a finite-state transducer that contains the relation of four word forms with the Quechua root *puñu*, 'to sleep' and their respective morphological analysis:[7]

$$
\begin{aligned}
R = \{ \quad & \text{puñu+1.SG} && : && \textit{puñuni} \\
& \text{puñu+2.SG} && : && \textit{puñunki} \\
& \text{puñu+PST+1.SG} && : && \textit{puñurqani} \\
& \text{puñu+PST+2.SG} && : && \textit{puñurqanki} \quad \}
\end{aligned}
$$

Note that the transducer contains an empty transition $\epsilon : \epsilon$ which makes the past suffix *-rqa* optional. The transducer in Fig. 2.4 can be applied in both directions:

1. Given *puñunki* as input, applied in 'upward' direction, it produces `puñu+2.SG` as output.
   This is the procedure for morphological analysis.
2. Given `puñu+PST+1.SG` as input, applied in 'downward' direction, it produces *puñurqani* as output.
   This is the procedure for generation.

[7] *puñuni* - 'I sleep', *puñunki* - 'you sleep', *puñurqani* - 'I slept', *puñurqanki* - 'you slept'

	variations	standard
progressive	*-chka/-sha/-sa/-sya*	*-chka*
genitive (after vowel)	*-p/-q/-h/-j*	*-p*
evidential (after vowel)	*-m/-n*	*-m*
additive	*-pis/-pas*	*-pas*
euphonic	*-ni/ñi*	*-ni*
1.&2. plural forms	*-chis/-chik/-chiq*	*-chik*
assistive	*-ysi/-schi/-scha*	*-ysi*
potential forms	*-swan/-chwan*	*-chwan*

TABLE 2.4: Suffix Variation and Normalization

2.3.2 Finite-State Analysis for Quechua

The original morphological analyzer was implemented in 2011 [Rios2011a], but has since undergone several changes. Both the original and the improved version are implemented in *xfst*[8].

While the original analyzer was a single finite-state transducer, the improved analysis tool consists of a chain of 5 cascaded transducers. The analyzer does not only split a given word form into morphemes, but additionally normalizes the surface form of the morphemes. Roots are mapped to their standardized form according to Cerrón-Palomino [1994], e.g. the word for girl (adolescent), *p'asña* in the standard, may also appear as *p'ashña, p'achña or pasña*, depending on the dialect. The analyzer rewrites all these variants to *p'asña*. Furthermore, it normalizes suffixes with variable dialectal forms, see Table 2.4.

Some of the suffixes from Table 2.4 are ambiguous in their non-standardized forms, e.g. the direct evidential suffix, written as *-n*, could also be a third person singular marker (verbal or nominal). In order to generate the normalized form of a word with a suffix *-n*, we need to know whether this particular *-n* is a person marker or an evidential suffix. Only in the latter case, *-n* needs to be rewritten as *-m* during normalization.

We have two normalizers in our pipeline: the first one handles text in 'regular' orthographies that show some minor dialectal variations. The second normalizer allows for more 'extreme' orthographies: for instance, both /k/ and /q/ (velar and postvelar stops) are pronounced as fricatives in certain positions ([x] and [χ]). In many texts, both are

[8] Xerox Finite State Tools, http://web.stanford.edu/~laurik/fsmbook/home.html

written as ⟨j⟩ (or sometimes ⟨h⟩) if pronounced as fricatives. This introduces new ambiguities: for instance, a root written as *sajsa* could be *saqsa* - 'certain variety of corn' or *saksa* - 'satisfied,full'. In order to avoid additional ambiguities resulting from an analysis with relaxed orthographic rules, the transducer with the additional orthographic rules handles only word forms that were not recognized by the standard normalizer.

As most Quechua texts contain Spanish words, we included two additional finite-state transducers that recognize Quechua words with Spanish roots.[9] The first one recognizes only word forms with correctly written Spanish roots, as in *vientota* - 'wind' (sp. *viento* and *-ta* - accusative), whereas the second transducer includes several rules that allow for an alternative spelling of the Spanish words, as in *huwista* - 'judge' (sp. *juez* and *-ta* -accusative). Furthermore, we implemented a guesser that attempts to split word forms into morphemes if the root is unknown. In order to prevent highly unlikely analyses, we restrict the guessing to roots of at least two syllables and with at least one Quechua suffix attached.

The five transducers are joined in a cascade: if the normalizer fails to analyze a word, the Spanish transducer is invoked. If this fails as well, the word is passed on to the second normalizer with relaxed orthography. If the word form has still no analysis, the second Spanish transducer with relaxed orthography attempts to find an analysis. Finally, if all transducers failed, the word is handed to the guesser. One of the Quechua texts used for evaluation, a story called *Hanaq pachaman wichaq wayna - El joven que se subió al cielo* [Lira 1990], contains relatively few words with Spanish roots, but in the other text, the last 72 sentences of the biography of Quechua native speaker Gregorio Condori Mamani, almost every sentence contains at least one word with a Spanish root. In this case, the Spanish transducer makes a considerable difference: coverage increases by roughly 22%, see Table 2.5. Furthermore, we tested the morphological analyzers on a third text, *Cancionero*, with an even more inconsistent spelling of Quechua words. The *Cancionero* contains religious (catholic) songs written in a 'Spanish' orthography, see the 'Spanish' example in Table 2.3. The restrictive Quechua and Spanish analyzers recognize only half of the word forms in this text, but the transducers with broader orthographic rules ('relax') increase the number of analyzed tokens to 98.43%, see Table 2.5.

[9]The lexicon contains all the Spanish lemmas, except function words, from FreeLing [Padró and Stanilovsky 2012]

	Joven	Gregorio	Cancionero
number of tokens:	1953	1024	1015
normalizer	97.86%	73.00%	42.56%
Spanish strict	0.64%	21.87%	15.86%
normalizer relax	-	-	34.88%
Spanish relax	-	0.30%	1.48%
guesser	1.02%	2.36%	3.65%
total coverage	99.52%	97.64%	98.43%
unknown words	0.48%	2.46%	1.58%

TABLE 2.5: Morphological Analysis Coverage

Figure 2.5 contains a simplified illustration of the finite-state transducer for the analysis (normalizer). The transducers *spanish* and *spanish-relax* are almost identical to *normalizer* and *normalizer-relax* respectively, but have different root types (Spanish nouns, verbs and some conjunctions). Table 2.6 contains a detailed listing of suffixes for the verbal slots (v1-v7), while Table 2.7 illustrates how the nominal slots (n1-n7) are filled. Table 2.8 contains the suffixes for the independent slots (i1-i7), but note that some possible combinations of suffixes are not consistent with the general order and have therefore been included as single tokens in the finite-state transducer (e.g. *-yari* and *-chuhina*). Furthermore, the independent suffix *-lla* is missing in Table 2.8, as there seems to be no restriction on its position: *-lla* may even be placed in between v1-v7 or n1-n7. Note that the only transition that has no ϵ alternative is the one from v5 to v6: a finite verb needs at least a person suffix (subject) to form a valid word[10], while for nouns, there is no such restriction: a bare nominal root is a grammatical Quechua word.

[10]There is one exception: in reduplicated forms, the first part is the bare root. Another special case are past tense suffixes *-rqa* and *-sqa*, the subject suffix can be dropped with 3.SG subjects. However, for the morphological analysis we consider those *-sqa* and *-rqa* to be portmanteau forms that include the 3.SG marker.

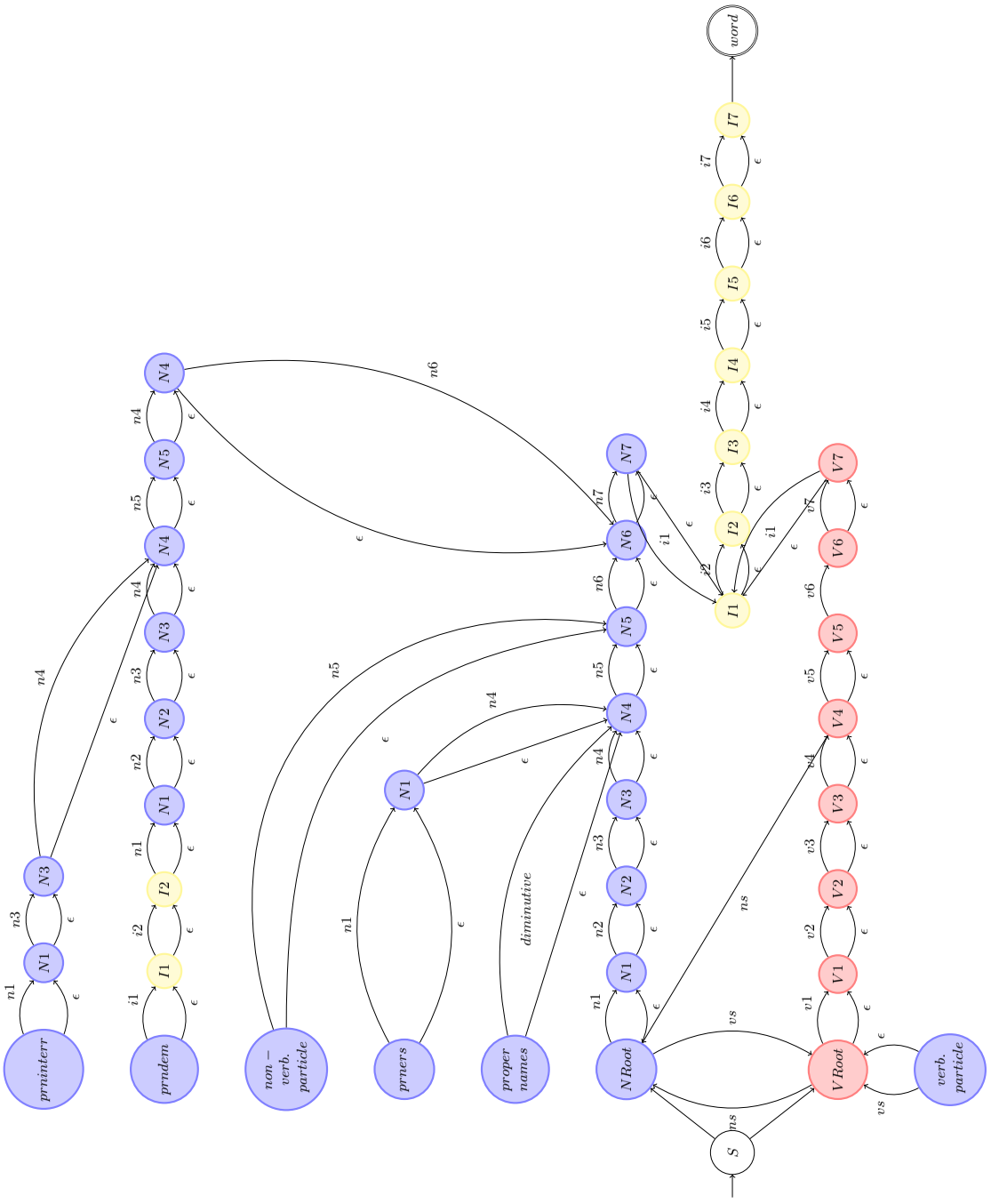


FIGURE 2.5: Finite-State Transducer For Quechua Morphology

2.4 Morphological Disambiguation and Text Normalization

Given the fact that a Quechua word form can contain more than one morphological ambiguity, the disambiguation has to be done in several steps. The simplest approach is to disambiguate each word form from 'left to right':

1. disambiguate the root (nominal vs. verbal)
2. disambiguate nominalizing and verbalizing suffixes
3. disambiguate verbal suffixes[11]
4. disambiguate independent suffixes

We use Wapiti [Lavergne et al. 2010], a toolkit for sequence labeling that includes an implementation of conditional random fields (CRF), in order to train 4 CRF models (one model for each step). We decided to use conditional random fields, as the task of morphology disambiguation is in many ways similar to part-of-speech tagging. There is an inter-dependency between the labels: the decision which label a given instance should receive depends to a certain extent on the labels of the previous n instances.

In the initial setup [Rios and Castro Mamani 2014], the training material comprised two Quechua texts that were analyzed with the xfst tools (see section 2.3) and then manually disambiguated: the biography of Quechua native speaker Gregorio Condori Mamani [Valderrama Fernandez and Escalante Gutierrez 1977], that contains about 2500 sentences, and some stories from a collection [Lira 1990], that amount to about 300 sentences. Once the annotation process of the treebanks was completed, we included two of the morphologically disambiguated texts into the training material (see chapter 3 for information about the treebank texts).

2.4.1 Model 1: Disambiguation of Ambiguous Roots

Many Quechua roots can be used nominally or verbally without derivation, see examples (5) and (6) on page 16. The disambiguation of roots can be regarded as part-of-speech

[11] There are no ambiguous sequences within the nominal suffixes, therefore the third step involves only verbal suffixes.

v1	v2	v3	v4	v5	v6	v7
	derivation	object	aspect	tense	person	mood
-lli	*-pu*	*-wa*	*-chka*	*-rqa*	subj	*-man* (POT)
-ri	*-mu*	*-su*	∼ *-sha*	*-sqa*	subj>obj	
-cha (VDIM)	*-ku*		∼ *-sya*	*-sqa*		
-paya	*-na* (RZPR)		∼ *-sa*			
-pa	*-chi*		∼ *-shia*			
-naya						
-rqu						
-yku						
-ysi						
∼ *-schi*						
∼ *-scha*						
-raya						
∼ *-laya*						
∼ *-nraya*						
-miya						
-nya						
-ykacha						
-tiya						
-rqari						
-yqari						
-rpari						
-ypari						
-ymana						
-pasa						
-rku						
-rpu						
-qi						
-naqa	Aymara loans					
-tata						
-thapi						
-kipa						
-kata						
-qa						
-t'a						

TABLE 2.6: Verbal Slots: Suffixes

n1	n2	n3	n4	n5	n6	n7
derivation		possesive	number		case	
-chika	*-sapa*	*-y*	*-kuna*	*-ntin*	*-ta*	*-kama*
∼ *-chikan*	*-yuq*	*-yki*		*-pura*	*-pa*	*-rayku*
-karay	*-nnaq*	*-n*		*-kama*	*-manta*	∼ *-layku*
-kankaray		*-nchik*		*-nka*	*-man*	*-wan*
-chaq		∼ *-nchis*		*-niq*	*-paq*	*-puwan*
-chachaq		∼ *-nchiq*			*-pi*	
-cha		*-yku*				
∼ *-scha*		*-ykichik*				
-niray		∼ *-ykichis*				
∼ *-niraq*		∼ *-ykichiq*				
-rikuq		*-nku*				
-ti						
-li		*-ykiku*		only in transitive nominalized verbs		
-liku		*-suyki*				
-lu		*-suykichik*				
-mpa		∼ *-suykichis*				
		∼ *-suykichiq*				
		-waykiku				
		-waychik				
		∼ *-waychis*				
		∼ *-waychiq*				
		-waykichik				
		∼ *-waykichis*				
		∼ *-waykichiq*				

Table 2.7: Nominal Slots: Suffixes

i1	i2	i3	i4	i5	i6	i7
-hina	*-puni*	*-pas*	*-taq*	*-chu*	*-mi*	*-iki*
∼ *-sina*		∼ *-pis*			*-si*	∼ *-riki*
-pacha		*-raq*			*-cha*	*-ya*
-chuhina		*ña*			∼ *-chi*	*-yá*
→ *-chu -hina*					*-qa*	∼ *-wá*
					-ri	*-yari*
					-suna	→ *-ya -ri*
					∼ *-sina*	*-chari*
					-má	→ *-cha -ri*
					-sá	
					-chá	

Table 2.8: Independent Slots: Suffixes

ALFS	alphabetic character
CARD	number
NRoot	nominal root
Part	particle
PrnDem	demonstrative pronoun
PrnInterr	interrogative pronoun
PrnPers	personal pronoun
VRoot	verbal root
SP	Spanish word
$	punctuation

TABLE 2.9: Root Tags

```
suwa   suwa[NRoot][=ladrón]

papanchikta   papa[NRoot][=patata][--]nchik[NPers][+1.Pl.Incl.Poss]
              [--]ta[Cas][+Acc]

tukunqa   tuku[NRoot][=lechuza][--]n[NPers][+3.Sg.Poss][--]qa[Amb][+Top]
tukunqa   tuku[VRoot][=acabar]][--]n[VPers][+3.Sg][--]qa[Amb][+Top]
tukunqa   tuku[VRoot][=acabar][--]nqa[VPers][+3.Sg.Fut]
```

FIGURE 2.6: Ambiguous Morphological Analysis for Example (7)

tagging with a very small tagset. Consider the following example (taken from a story in Lira [1990]):

(7) *..suwa papa -nchik -ta tuku -nqa..*
thief potato -1.PL.INCL.POSS -ACC end -3.SG.FUT
'[..] the thief will take all our potatoes [..]
(lit. 'the thief will end our potatoes')

possible lemmas	case	possible root tags	possible morph tags
suwa	lc	NRoot	-
papa	lc	NRoot	1.PL.INCL.POSS, ACC
tuku	lc	NRoot, VRoot	3.SG.POSS, TOP, 3.SG, 3.SG.FUT

TABLE 2.10: Features for Disambiguation with Wapiti for Example (7)

The root *tuku-* 'to end' is ambiguous: *tuku-* can also be a nominal root with the meaning 'owl'. Furthermore, the sequence *-nqa* is ambiguous, apart from the 3rd singular future form, it could be a combination of *-n*, '3rd singular subject' or '3rd singular possessive', and *-qa*, 'topic', see Fig. 2.6 with the output of the xfst analyzer for this example. In a first step, the type of the root has to be determined, the ambiguity of *-nqa* is only

relevant if the root is verbal and will be dealt with later. In order to disambiguate the root with Wapiti, every token needs to be converted into a set of features (an instance) extracted from the xfst output, see Table 2.10. The words *suwa* and *papanchikta* are not ambiguous[12] and therefore have only one possible root tag, whereas *tukunqa* has two possible root tags: VRoot and NRoot. Model 1 will assign one of them as class label, considering the features and the context of the given token. Wapiti allows pre-labeled input data, therefore, we can already set the label of the unambiguous words *suwa* and *papanchikta*. Note that the instances do not contain the full word form; due to the small size of our training corpus, using full word forms leads to increased data sparseness and impairs the results.

2.4.2 Model 2: Disambiguation of Nominalizing and Verbalizing Suffixes

Even after the disambiguation of the root type, the final word form can still be either nominal or verbal, as certain nominalizing and verbalizing suffixes are homophonous with verbal or nominal morphemes. Consider the following examples:

(8) *wasi -cha -y*
house -FACT(VS) -INF(NS)/-2.SG.IMP
'to build a house' or 'build a house!'

wasi -cha -y
house -DIM -1.SG.POSS
'my small house, cottage'

(9) *riku -sqa -yki*
see -PERF(NS) -2.SG.POSS
'the one you saw, your seeing'

riku -sqayki
see -1.SG>2.SG.FUT
'I will see you'

The suffix *-cha* attached to a nominal root can be either a diminutive or a factitive suffix ('make'): With the diminutive, the resulting word form is still a noun, whereas the factitive suffix produces a verb. In total, model 2 handles eight different cases of ambiguous verbalizing/nominalizing vs. verbal/nominal suffixes. The features in models 2-4 are essentially the same as those in model 1 (see Table 2.10), but of course the root type is no longer ambiguous, consequently there is only one root tag. With models 2-4, we classify only words that exhibit a verbalizing/nominalizing vs. verbal/nominal

[12]The root *suwa* is actually ambiguous, since it can be nominal ('thief') or verbal ('to steal'). The bare root without suffixes can however only be the noun.

ambiguity, whereas words that are unambiguous for this particular model receive a dummy label ('none').

2.4.3 Model 3: Disambiguation of Verbal Morphology

In the next step, we disambiguate six possible ambiguities in verb forms. One of the ambiguities in question is the sequence *-nqa* from example (7): after applying model 1, we know that the root *tuku* in *tukunqa* is verbal, but *-nqa* can still be either the 3rd singular future form or a combination of 3rd singular present and topic marker, see example (10). Other ambiguities of this type involve *-sun*, which can be either the imperative or future form of the first plural inclusive, as well as the sequence *-sqaykiku*, which can be either the indirect past or future form of the first plural exclusive acting on a 2nd singular person.

(10) *tuku -nqa*
end -3.SG.FUT
'he will end'

tuku -n -qa
end -3.SG -TOP
'he ends'

(11) *llamk'a -sun*
work -1.PL.INCL.FUT
'we will work'

llamk'a -sun
work -1.PL.INCL.IMP
'let's work'

(12) *qhawa -sqaykiku*
look -1.PL.EXCL.>2.SG
'we (excl.) watch you'

qhawa -sqa -ykiku
look -IPST -1.PL.EXCL
'we (excl.) watched [they say]'

2.4.4 Model 4: Disambiguation of Independent Suffixes

Model 4 disambiguates ambiguities that concern independent suffixes. None of these potential ambiguities occur in all dialects and orthographies, but all of them concern the normalization and are therefore important. There are 3 types of ambiguities that relate to independent suffixes:

The most common case involves the suffix *-n*, when the word form is nominal and *-n* follows a vowel: in this case, *-n* can be the 3rd singular possessive, or it can be the allomorph of the evidential suffix *-mi*. The latter is written as *-m* in the standard orthography, as well as in texts written in Ayacucho Quechua, but occurs as *-n* in many

texts written in Cuzco and Bolivian Quechua, see example (13). A further ambiguity that occurs only in Cuzco and Bolivian Quechua concerns the sequence *-pis*: *-pis* can be the additive suffix (in Ayacucho Quechua always *-pas*) or a combination of the locative suffix *-pi* and the evidential suffix *-s*, see example (14). The third ambiguity of this type concerns Spanish words that end in *-s*: in this case, *-s* can be an evidential suffix, but it can also be the Spanish plural[13], see example (15).

In addition to the 3 cases mentioned above, there is one more possible ambiguity that involves the direct evidential suffix *-n*: the sequence *-rqakun* can be segmented as *-rqa -ku -n* (-RPTN -RFLX -3.SG) or *-rqaku -n* (-3.PL.PST -DIRE).[14] In the second case, the normalized form should be *-rqakum*. This sequence occurs neither in our small training corpus nor in the test set. For this reason, the system cannot resolve this ambiguity at the moment, but we plan to train our models in the future with more material, so that we can include this case into the disambiguation process.

(13) *wasi -n*
house -DIRE
'house'

wasi -n
house -3.SG.POSS
'his house'

(14) *chay -pis*
this -ADD
'also this'

chay -pi -s
this -LOC -INDE
'there [they say]'

(15) *derechu -s*
right -INDE
'right [they say]'

derechu -s
right -PL
'rights'

2.4.5 Performance of the Four Models

Table 2.11 contains the results obtained with all four models trained and tested on different subsets out of these texts:

GREG Autobiography Gregorio Condori Mamani [Valderrama Fernandez and Escalante Gutierrez 1977], minus the last 70 sentences

[13] In some Bolivian dialects *-s* is also used on native roots as plural suffix, see the Bolivian word *p'achasta* (normalized *p'achakunata*) in Table 2.3 on page 17.

[14] *-ku* in the latter analysis could also be separated as plural marker, the morphemes would then be (-PST -PL -DIRE). Nevertheless, we treat the person and plural marker as a unit for practical reasons.

GREG70 Last 70 sentences of autobiography GREG

LIRA *Cuentos del Alto Urubamba*, [Lira 1990], minus the story *El joven que se subió al cielo*

JOVEN *El joven que se subió al cielo*, one of the stories in LIRA

AHK Festschrift 40th anniversary of the Peruvian-German chamber of commerce and industry (322 sentences), translated by C. Morante Luna, part of the SQUOIA treebank

INFO *La papa y el cambio climático* - 'potatoes and climate change', inforesources 2008 (development aid, 456 sentences), translated by I. Álvarez Ccoscco and C. Morante Luna, part of the SQUOIA treebank[15]

GREG and LIRA are relatively similar as to narration style and the occurrence of certain forms: both texts contain stories, either personal or legends. Both AHK and INFO, on the other hand, are texts that were translated from Spanish: they do not contain stories, but rather technical descriptions of given political, economical or ecological processes. Style, content and domain differ considerably between GREG and LIRA on one side, and AHK and INFO on the other side.

The results show a clear correlation as to the text genre and domain of training and test set: while the addition of more training material does not have a huge impact on the test sets derived from GREG and LIRA, adding INFO to the original training set improves the results on AHK considerably. Furthermore, the inclusion of AHK into the training material increases the correctly disambiguated word forms in INFO.

This correlation becomes especially evident in the results for model 2, which includes the disambiguation of the suffix *-sqa* (see section 2.4.2). This suffix can be a nominalizing suffix (perfect marker) or a tense suffix that marks the so-called narrative past.[16] In both AHK and INFO, *-sqa* often occurs within *nisqa* 'called, said', which provides a special way of including foreign words in Quechua: instead of attaching suffixes directly to the foreign term, *nisqa* is used in order to bear the necessary suffixes, see example (16): the foreign term *Tratado Bilateral* ('bilateral contract') is the object of the verb *hunt'a-*

[15] See Chapter 3 for details about the treebank texts.

[16] a special past form that indicates indirect evidence, i.e. the speaker did not witness the event himself. This form is especially frequent in story telling, hence the name 'narrative past'.

('to fulfill'), and thus needs an accusative marker. Instead of attaching the accusative suffix directly to *Bilateral*, *nisqa* follows the foreign phrase and bears the corresponding suffixes.

(16) *Qullqi churana amachanapaq Tratado Bilateral* ***nisqata*** *hunt'akuyninqa [..]*

Qullqi chura -na amacha -na -paq Tratado Bilateral ***ni -sqa -ta***
money put -OBL protect -OBL -BEN contract bilateral say -PERF -ACC
hun'ta -ku -y -nin -qa
fulfill -RFLX -INF -3.SG.POSS -TOP
'As for the fulfillment of the bilateral contract for the protection of investments [..]' [translation of AHK]

This special use of *nisqa* almost never occurs in GREG and LIRA. In LIRA the narrative past form *-sqa* occurs with about the same frequency as the perfect form, but with the root *ni-* ('to say'), *-sqa* occurs almost exclusively as the finite narrative past form:

(17) *Willaptintaqsi tayta mamanqa* ***nisqa****: -Kunan pampachasqayki suwachimusqayki-manta.*

Willa -pti -n -taq -si tayta mama -n -qa ***ni -sqa***
tell -DS -3.SG.POSS -ADD -INDE father mother -3.SG.POSS -TOP say -IPST
-ø: Kunan pampacha -sqayki suwa -chi -mu -sqa -yki
-3.SG now forgive -1.SG>2.SG.FUT steal -CAUS -DIR -PERF -2.SG.POSS
-manta
-ABL
'When he told [them], his father and mother **said**: -Now I forgive you for letting them rob you.' [Lira 1990]

Table 2.11 shows that for the test set GREG70 and JOVEN, which are both parts of GREG and LIRA, the performance of model 2 drops when the training material from the treebanks is included. On the other hand, both AHK and INFO as test set benefit substantially from the inclusion of the respective other technical text.

For the disambiguation of the evidential suffix *-n/-m* and the additive suffix *-pis/-pas* the system includes two parameters that can be set manually: if the text to be disambiguated is not written in an orthography where *-n* can be the form of the evidential *-m*, the system will discard all analyses for the suffix *-n* as evidential and instead treat it always as 3rd person marker. Likewise, the need to process the additive suffix *-pis* can be set with a

		training set	Greg70	Joven	Info	AHK
model 1	root tag	Greg, Lira	95.35	**85.71**	85.84	91.97
		Greg, Lira, Info	**97.67**	84.52	–	**92.17**
		Greg, Lira, AHK	**97.67**	84.52	**89.21**	–
		Greg, Lira, Info, AHK	95.35	83.33	–	–
		baseline	65.12	72.62		
model 2	NS/VS	Greg, Lira	**97.44**	**87.88**	84.54	73.74
		Greg, Lira, Info	92.31	84.34	–	**86.82**
		Greg, Lira, AHK	92.31	84.34	**93.18**	–
		Greg, Lira, Info, AHK	94.87	79.39	–	–
		baseline	80.49	17.47		
model 3	verbal s.	Greg, Lira	85.71	66.67	82.93	54.17
		Greg, Lira, Info	85.71	66.67	–	54.17
		Greg, Lira, AHK	85.71	66.67	**87.18**	–
		Greg, Lira, Info, AHK	85.71	66.67	–	–
		baseline	**88.89**	**75.00**		
model 4	independent s.	Greg, Lira	85.37	86.11	100	100
		Greg, Lira, Info	82.93	83.33	–	100
		Greg, Lira, AHK	82.93	83.33	100	–
		Greg, Lira, Info, AHK	79.49	**86.49**	–	–
		baseline	64.10	50.00		

Table 2.11: Evaluation: Precision of the Morphological Disambiguation Steps

5-fold cross-validation on all texts		
model 1	root tag	92.64
model 2	NS/VS	84.81
model 3	verbal suffixes	77.33
model 4	independent suffixes	82.63

Table 2.12: 5-Fold Cross-Validation

parameter, in case the text is not written in a dialect where the additive occurs as *-pis*, this analysis will be discarded and all occurrences of *-pis* will be treated as combination of locative *-pi* and indirect evidential *-s*. Since in both treebank texts the direct evidential is written as *-m* and the additive suffix as *-pas*, there is no need to disambiguate these suffixes, hence the 100% correct forms on Info and AHK for model 4 in Table 2.11.

Table 2.12 contains the results of a 5-fold cross-validation on all four texts combined. Note that in the case of model 4, the disambiguation of the independent suffixes, we cannot set the parameters for the disambiguation of the evidential *-m/-n* and the additive suffix *-pis/-pas* since the test sets contain an arbitrary subset of all texts. The results for model 4 are thus lower than they would be normally.

2.4.6 Evaluation

We used the same test sets as for the evaluation of the morphological analysis in section 2.3: The last 72 sentences from the autobiography of Gregorio Condori Mamani [Valderrama Fernandez and Escalante Gutierrez 1977], and the Andean story *El joven que se subió al cielo* from Lira [1990] with about 250 sentences. Both test texts were excluded from the training set.

Table 2.11 illustrates the percentage of correctly disambiguated words with the particular ambiguity for each step. Note that there were only a handful of test cases for model 3 (verbal suffixes) in both texts, therefore, the results for this step might not be accurate. Furthermore, the number of instances extracted from the training material for model 3 is smaller than for the other models, as these types of ambiguities are relatively rare. For the normalization, errors in model 3 do not affect the outcome, as these ambiguities have no effect on the surface forms in the standard orthography. Considering for instance example (10), *-nqa* will be *-nqa* in the standard, irrespective of whether the analysis is *-n -qa* or *-nqa*.

Table 2.13 contains the evaluation of the disambiguation on the complete texts. Although the percentage of tokens with a wrong morphological analysis is almost the same in both texts, the total number of correctly analyzed words is lower in the biography. This is due to the fact that this text contains many words with Spanish roots, often with 'quechuized' spelling. Some of these words were not recognized by the xfst analyzer and were therefore not normalized.

The baseline for both Table 2.11 and 2.13 was calculated based on the frequencies of the forms in the training material: the baseline shows the results that we obtain if we disambiguate the test texts by choosing always the most frequent class in every decision. The difference with this approach, as opposed to the Wapiti models, is that we do not consider any context information for the baseline. Table 2.11 shows that Wapiti outperforms the baseline in every step except for model 3, for which the training instances are too sparse. There is a considerable difference in the baseline for the two test texts (see Table 2.13): on the biography, the baseline is much higher, since that the largest part of the training material is part of the same book. For this reason, the probability distribution of the individual classes in this test text correlates better with

the frequencies calculated from the training material. This becomes especially evident once again in the disambiguation of the suffix *-sqa* (see section 2.4.2), which can be a nominalizing suffix (participle/perfect marker) or a tense suffix that marks the so-called narrative past. In the biography, and thus the largest part of training material, the latter form is relatively rare, most uses of *-sqa* are perfect forms. In *El joven que subió al cielo*, on the other hand, the narrative past form occurs with about the same frequency as the perfect form, but as the perfect was more frequent in the training corpus, all the narrative past forms are labeled as perfect forms with the baseline approach.

On the test set that is similar to the training material, the conditional random fields improve the disambiguation only slightly compared to the baseline (+2%), however, the effect they have on the results for a test set from a different text is considerable: >10%. Table 2.13 also contains the results obtained with the RFTagger [Schmid and Laws 2008] and Morfette [Chrupala et al. 2008] for comparison.[17] The main difference between our approach and the morphological taggers is that the latter analyze and label the complete word form at once, whereas with our approach, we disambiguate and normalize each word in several steps, proceeding from left to right. The tagset used by the morphological taggers is thus much more fine-grained, as each tag contains the morphology of the whole word. The results show clearly that our training corpus is too small to achieve satisfactory results with morphological tagging. It has to be noted as well that the RFTagger makes basic morphological assumptions about word forms that suit inflectional languages, such as German, but not agglutinative languages. The data format of the RFTagger relies on a set of grammatical features in a given word class that have exactly one value at a time, for instance, a feature in German nouns is *case* with the possible values *nominative, accusative, dative or genitive.* In Quechua, however, nouns might have more than one case suffix, and the issue is even more difficult with verbs: for instance, the derivational suffixes in verbal slot 1 (v1) from Table 2.6 on page 27 are optional, but a word might as well have several of these suffixes combined. We can therefore not simply assume a single category for those suffixes. Furthermore, even the basic distinction of verbs and nouns is not as straightforward as in other languages: nominalized verbs can bear case suffixes and are thus clearly nominal, while at the same time having clearly verbal features, such as object or aspect markers. Of course we can

[17] Both taggers were trained with default settings; the best results for the RFTagger were obtained on bigrams, as indicated in Table 2.13.

	El joven que subió al cielo		*Gregorio C. Mamani*	
total sentences:	258		72	
total token	1865		1015	
punctuation marks:	567		171	
xfst failures:	9	0.48%	25	2.46%
total word forms	1298		844	
correct analysis:	1252	**96.46%**	789	**93.48%**
wrong analysis:	33	2.54%	17	2.01%
guessed, no analysis in gold:	4	0.31%	6	0.71%
ambiguous words:	282	21.73%	127	15.05%
still ambiguous:	0		7	5.51%
correct of ambig.:	249	88.30%	103	81.10%
wrong of ambig.:	33	11.70%	17	13.39%
morphological tagging (tag whole word form):				
RFTagger (bigrams):		65.49%		72.21%
Morfette:		65.10%		78.32%
baseline (left to right, most frequent morphemes, no context considered):		85.98%		91.0%

TABLE 2.13: Evaluation: Disambiguated Texts

just treat nominalized verbs as a part-of-speech on their own with nominal and verbal features, which was the approach we used in our experiments. However, this leaves us with a very large set of features for the tags, and given the small training corpus, the results are poor.

As mentioned above, not all ambiguities are relevant for the normalization. In fact, many morphological ambiguities are irrelevant for the conversion to the standard orthography, therefore, the number of correctly normalized forms is higher than the proportion of correctly disambiguated words from Table 2.13. In the text *El joven que subió al cielo*, the percentage of correctly normalized words amounts to 99.61%, whereas for the biography of Gregorio Condori Mamani, we achieve 98.93%.

Table 2.14 contains the end result of the 5-fold cross-validation from Table 2.12 on page 35: on average, 95.49% of the words ended up with the correct morphological analysis, which is in accordance with the results of the smaller evaluation on only two texts in Table 2.13.

fold 1	fold 2	fold 3	fold 4	fold 5	Average
94.96	95.67	95.68	95.54	95.59	**95.49**

TABLE 2.14: Disambiguated Texts: 5-Fold Cross-Validation on GREG, LIRA, INFO and AHK

2.5 Spell Checking

Spell checking is an important tool for writing, and almost every text processing system has a device to avoid misspelled words.[18] Basically, the process of spelling correction consists of two tasks: in a first step, the spell checker decides whether a given word form is correct. If this is not the case, it has to find correct word forms close to the given word and suggest them as possible corrections.

Spell checking methods developed for languages like English usually rely on complete lists of word forms, a condition that cannot be met for morphologically complex languages such as Quechua.

Each nominal or verbal root in Quechua may be used in thousands of possible word forms, therefore the compilation of fully fledged word lists is not feasible. A better approach to capture the morphological structures of agglutinative languages are finite-state networks. Spell checkers that work with finite-state networks have been described, amongst others, for Turkish [Oflazer 1996], Finnish [Pirinen and Lindén 2010] and Basque [Alegria et al. 2002].

The pipeline for the analysis of Quechua word forms described in the previous sections can easily be adapted for spell checking. The setup is similar to the normalization, we use several finite-state transducers, implemented however in foma [Hulden 2009b] instead of *xfst* since foma includes the option MED search: with this option, foma finds strings in a network that are within a given minimum edit distance (hence MED) from the input string.[19]

1. strict analyzer: decide whether word form is correct

[18] Parts of this section are based on Rios [2011b] and Castro Mamani and Rios Gonzales [2014].

[19] The minimum edit distance between two string is the smallest number of basic edit operations (deletion, insertion and substitution of characters) that is necessary to convert one string into the other.

2. if not correct: try to rewrite word to its normalized form with 4 cascaded transducers:
   (a) normalizer: analyzer that allows different orthographies and certain dialectal variations, but rewrites all morphemes to the official standard as defined by R. Cerrón-Palomino [1994]. Example: *puriqtin* → *puriptin*
   (b) spanish: same as normalizer, but contains the Spanish lexicon from FreeLing; recognizes words with Spanish roots. Example: *tocashan* → *tocachkan*
   (c) normalizer-relax: same as (a), but allows for more extreme orthographic variations.
   Example: *huasiqui* → *wasiyki*
   (d) spanish-relax: same as (b), but allows for more extreme orthographic variations.
   Example: *iglisiya* → *iglesia*
3. if normalization failed: use minimum edit distance (MED) search on strict analyzer

In the first version of the spell checker, we used only the strict analyzer in combination with foma's own MED search [Hulden 2009a] to find suitable suggestions for misspelled words. However, it soon became evident that this procedure did not produce satisfactory results, as it relies only on edit distance, but the word closest in edit distance is not necessarily the intended word. Therefore, including some knowledge about dialectal and orthographic variation and thus anticipating frequent mistakes improves the spell checking considerably.

Consider the following example with the misspelled word *rimashan*:

(18) *rima -sha -n*
speak -PROG -3.SG
'he is speaking'
normalized: *rima -chka -n*

We can be fairly certain that the intended word form was *rimachkan*, as *-sha* is a dialectal variation of the progressive suffix *-chka*. Nevertheless, if we use edit distance as the only error metric, other word forms are closer and we get these suggestions on top of the list: *rimaspan, rimasqan* and *rimachan*, as for all three, edit distance to *rimashan* is only 1 (substitute *h* with *p* or *q*, or substitute *s* with *c*), whereas the edit distance to *rimachkan*

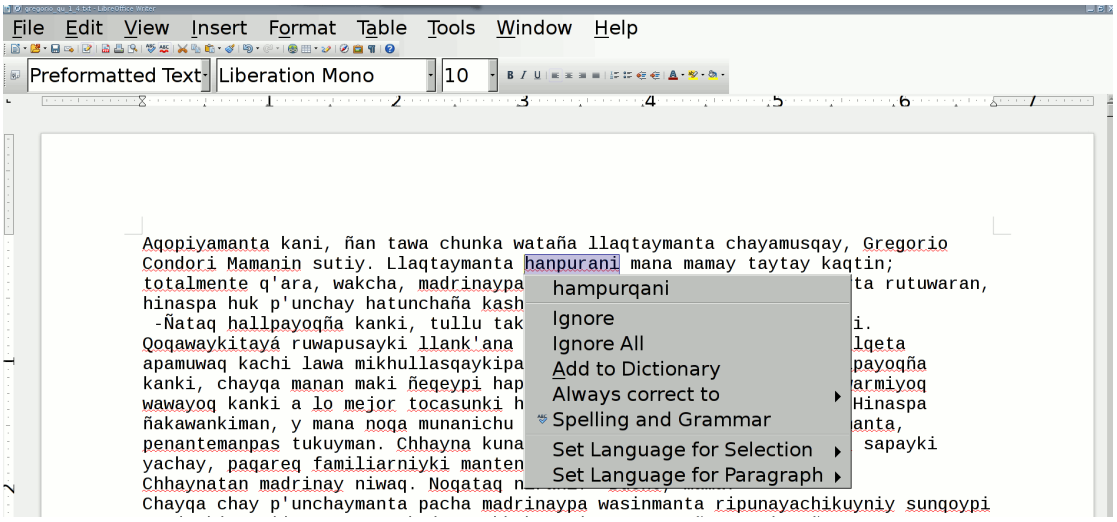


FIGURE 2.7: Spell Check Plugin in LibreOffice Writer

is 2 (insert *k*, substitute *s* with *c*). With the normalization pipeline, however, we simply rewrite the progressive suffix *-sha* to the normalized form *-chka* and thus avoid unwanted suggestions.

The four cascaded finite-state transducers (2a-d) result in a relatively large network, and since we have to load this network into the memory with every call (every misspelled word), spell checking was too slow to provide a comfortable writing assistance. For this reason, we implemented the spell checker in a server-client architecture: the server loads all transducers at start-up and then waits for words sent by the client.

Our back-end spell checker has been included in plugins for OpenOffice and LibreOffice, as well as for Microsoft Office by Richard Castro Mamani from the group *hinantin* at the *Universidad Nacional San Antonio Abad del Cusco* (UNSAAC), see Fig. 2.7 with the spell checker in LibreOffice.[20]

2.6 Summary

This chapter laid the foundation for automatic processing of Quechua texts on word level. Quechua texts come in a large variety of different spellings: there are several dialectal distinctions on the lexical, phonological and morphotactical level within the Southern Quechua varieties. Furthermore, there has been a long ongoing debate about

[20]The plugin is available from: https://github.com/hinantin/LibreOfficePlugin.

the correct orthography for Southern Quechua, and even though the Peruvian Ministry of Education declared the standard defined by Cerrón-Palomino [1994] official, many texts are still written in a different spelling. In this chapter, we presented our approach to automatically parse the word forms with a set of finite-state transducers and how we use machine learning to disambiguate words with more than one possible analysis. Once a word form has been disambiguated, we can simply rewrite the morphemes in their standardized form. This automatic normalization is needed for any further processing of Quechua text: the language model for the translation system (see chapter 5) was trained on automatically normalized text, and also the first step in the annotation of the SQUOIA treebanks (see chapter 3) consists of a morphological analysis and disambiguation of the texts. Furthermore, we showed that, even though the morphology tools were not developed with spell checking in mind, we were able to adapt them to this application with little effort.

Chapter 3

Quechua Treebank

3.1 Introduction

An important part of the SQUOIA project was the creation of a multilingual treebank in the three languages of the machine translation systems Spanish-German and Spanish-Cuzco Quechua,[1] since the original idea was to use the treebanks in order to improve the hybrid machine translation systems that constitute the main part of the SQUOIA project.[2] Treebanks can be regarded as a special type of annotated corpora: a collection of texts annotated with grammatical information beyond the part-of-speech level. The syntactic structure of the sentences in a treebank is mapped to a graph that is usually displayed as an inverted tree, where the top node represents the root and the tokens represent the leaves [Adesam 2012:28]. The form of the trees and the syntactic information depend on the grammar formalism behind the annotation scheme.

A widely used approach in the syntactic annotation of texts are dependency grammars, where the notion of *dependency* is based on the idea that the syntactic structure of a sentence consists of binary asymmetrical relations between the words of the sentence [Nivre 2005:3]. In other words, a dependency relation is the relation between a head and a dependent. While this general definition is true for all varieties of dependency grammars,

[1] Even though the SQUOIA treebank is multilingual, this chapter belongs to the part about monolingual resources since it deals only with the Quechua part of the treebank. The German and Spanish counterparts are annotated according to a different syntactic formalism and will not be discussed here.

[2] Sections 5.3 and 5.8 from Chapter 5 illustrate how we rely on treebanks for certain sub-tasks during the translation process.

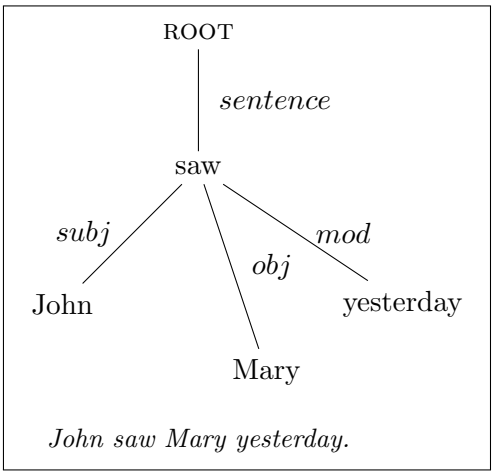


FIGURE 3.1: English Dependency Tree Example

they differ according to the exact criteria as to what defines a head or a dependent. The head will usually determine the syntactic and semantic category of a given structure, furthermore, the form of the dependent is defined by the head. Typical dependency relations are thus head-complement and head-modifier structures. However, for many other syntactic constructions, especially those involving function words, such as articles, complementizers, auxiliary verbs or adpositions, the relation is not as straightforward. As a consequence, the treatment of these structures differs considerably between different varieties of dependency grammars [Nivre 2005:4-6].

In contrast to the more common phrase structure annotations that group words into phrases defined by structural categories, such as *verbal phrase* or *noun phrase*, a dependency annotation represents head-dependent relations between words, labeled with functional categories, such as *subject* or *object* [Kübler et al. 2009:2], see Fig. 3.1 with a simple English dependency tree.

A dependency tree as defined in Kübler et al. [2009:12] is a labeled directed graph $G = (V, A)$ that consists of nodes (V), arcs (A) and a set of possible relation types $R = r_1..r_n$, such that for a sentence $S = w_0w_1...w_n$ and a label set R the following holds:

1. $V \subseteq w_0, w_1, ..w_n$
2. $A \subseteq V \times R \times V$
3. if $(w_i, r, w_j) \in A$ then $(w_i, r', w_j) \notin A$ for all $r' \neq r$

A dependency graph G is thus a set of labeled dependency relations between the words of S such that all words are connected and there is exactly one relation between two words [Kübler et al. 2009:12]. The third restriction does not apply to multi-layered treebank annotations, where words might be connected by more than one arc, each representing a different layer of annotation (e.g. syntactic vs. semantic).

The following sections describe the corpus and the concrete dependency scheme that we designed to annotate the Quechua texts in the project SQUOIA.

3.2 Corpus

The goal in the SQUOIA project was to build parallel treebanks in all three languages of the project: Spanish, (Cuzco) Quechua and German. Naturally, parallel texts that are available in these three languages are rare, therefore the majority of the corpus consists of parallel Spanish-German texts, while the Quechua counterpart was translated by several native speakers in Peru as part of the SQUOIA project, with one exception: the biography of Gregorio Condori Mamani [Valderrama Fernandez and Escalante Gutierrez 1977] is the only text in our treebank that was originally written in Quechua and only later translated to Spanish and German.[3] The treebank contains two chapters (∼ 500 sentences) of Gregorio Mamani's story.

For the remaining texts for the treebank, we selected reports on agriculture, development aid, economy, education, media and culture. All of them are freely available in Spanish and German and the translations are of good quality. Furthermore, all texts are somehow related to Peru or at least Latin America:[4]

- annual report 2009 of the Deutsche Welle Academy about *Development and the Media* [5]

[3] The Spanish and German translation of the biography are available as books and have been translated outside the SQUOIA project.

[4] The syntactic and morphological annotation has been completed so far only for the first 3 texts. However, all texts have been translated and are available in their Quechua version from the SQUOIA website at `https://github.com/ariosquoia/squoia/tree/master/treebanks/texts/quz`

[5] `http://www.dw.de/`

- *La papa y el cambio climático* - 'potatoes and climate change', inforesources 2008 (development aid)[6]
- Festschrift 40th anniversary of the Peruvian-German chamber of commerce and industry[7]
- *La revolución ganadera: ¿Una oportunidad para los productores pobres?* - 'The Livestock Revolution: An Opportunity for Poor Farmers?, inforesources 2007 [8]
- Memoria 2009, Peruvian-German chamber of commerce and industry[9]
- strategy paper of the Swiss Agency for Development and Cooperation on the co-operation with Peru [10]
- annual report 2008 of a private foundation dedicated to education[11]
- annual report 2010 of the International Monetary Fund (IMF)[12]
- anniversary publication of the Austrian Institute for Latin America[13]

We asked Quechua native speakers to translate these texts. The translation of these texts into Quechua is difficult, as they contain many terms or concepts that are not part of the traditional Andean context and thus have no direct correspondence in Quechua. Generally, there are three possible solutions to this problem:

1. the translator creates a term in Quechua that describes the same meaning.
   - *bautizar* - 'to baptise' > *suti churay* - 'to put a name'
   - *hidropesía* - 'hydropsy' > *punkillikuy unquy* - 'the self-swelling-disease'
2. loan words: the translator uses the Spanish root with Quechua morphology.

   ..sirtificadu -kuna -ta qu -rqa -n.
   certificate -PL -ACC give -PST -3.SG
   '.. he/she gave the certificates.' (Spanish: *certificado*)
3. foreign words: the translator uses the Spanish word(s) and cites with *nisqa* - 'said,called'; *nisqa* then bears the suffixes that correspond to the foreign word(s).

[6] http://www.inforesources.ch/pdf/focus08_1_s.pdf
[7] http://www.camara-alemana.org.pe/Publicaciones/MIGEdiciones/2010MEMORIA2009.pdf
[8] http://www.inforesources.ch/pdf/focus07_1_s.pdf
[9] http://www.camara-alemana.org.pe/Publicaciones/MIGEdiciones/2010MEMORIA-JAHRESBERICHT2009x.pdf
[10] *Schweizerische Kooperationsstrategie für Peru 2009-2011*, http://www.deza.admin.ch/de/Home/Dokumentation/Publikationen?page=107
[11] http://www.fundeducation.org/
[12] http://www.imf.org/external/pubs/ft/ar/
[13] http://www.lai.at/

Parlamento alemán ni -sqa -p uma -lli -q -nin
parliament german speak -PERF -GEN head -AUTOTRS -AG -3.SG.POSS
'the leader of the German parliament' (Spanish: *parlamento alemán*)

There is a clear difference between loan and foreign words: the former are typically written according to the Quechua pronunciation and receive the same treatment as native Quechua roots, i.e. they can bear suffixes. Foreign words, on the other hand, keep the original Spanish spelling and do not bear suffixes, but instead are 'cited' with *nisqa* - 'called, said'; this element then bears all the corresponding suffixes.
For the treebank texts, the goal was to have as few foreign words as possible in the Quechua texts. Nevertheless, there were cases where it would have been rather confusing to invent a native construction instead of using the Spanish term with *nisqa*. Every individual case needs to be considered carefully, which makes the translation process particularly difficult.

3.3 Quechua Dependency Annotation Scheme

Contrary to most other treebanks, the basic units in the Quechua treebank are morpheme groups instead of complete word forms.[14] The reason behind this approach is that many syntactic and semantic features that in languages such as English or Spanish are expressed through function words correspond to Quechua suffixes. In order to allow these suffixes to receive their own dependency label, we decided to split the words into morpheme groups, an idea based on the description of the Turkish METU-Sabanci treebank [Atalay et al. 2003, Eryiğit 2007].

The following section illustrates some basic features and special cases in the Quechua annotation scheme. First of all, we do not include punctuation marks directly in our dependency trees, instead, we introduce a non-terminal node, a virtual root (VROOT). Every punctuation mark depends directly on this virtual root as *punc*, whereas the dependency tree depends as *sntc* (sentence) on VROOT.

[14] Parts of this chapter are based on Rios and Göhring [2012] and Rios [2014].

3.3.1 Case Suffixes

We consider case markers and postpositions as equivalent to prepositions in languages such as English (e.g. Quechua instrumental case *-wan* corresponds to English 'by, with'). In accordance with the Stanford Dependency scheme [de Marneffe and Manning 2008] we treat case suffixes as the head of the noun they modify.

3.3.2 Elision of Copula

The copula *ka-* - 'to be' is often elided in third person contexts (see examples (25) and (26)). Apart from equational clauses, these include third person singular habitual past and obligative forms. In this case, we insert a dummy element (*KAN*), as the verbless clause would lack a head otherwise.

3.3.3 Coordination

Coordinations are headless constructions by nature, therefore we arbitrarily annotate the last element as head of the preceding coordinated elements. For a head-final language like Quechua, it makes more sense to treat the last element as head instead of the first, as this is in accordance with other constructions. Coordinations can be expressed through a limited set of coordinative particles[15] but also through suffixes, or both. In coordinations involving connective suffixes usually every element is morphologically marked for coordination. See Fig. 3.2 with the simplified annotation of the following examples:[16]

(19) *Mana -m uywa -y -pas ni chakra -y -pas ka -n*
Not -DIRE animal -1.SG.POSS -ADD nor field -1.SG.POSS -ADD be -3.SG
-chu.
-NEG
'I don't have animals nor field.'
(lit. 'Neither my animal nor my field exists.')

[15] e.g. *icha* - 'or' and postposition *ima* - 'also'; additionally, combinations of demonstrative pronouns with case or so-called independent suffixes may serve as clause linkers. Furthermore, Spanish borrowings like *ni* -'nor, neither' are frequently used in texts.

[16] For a list of edge labels, see Abbreviations on page xvi.

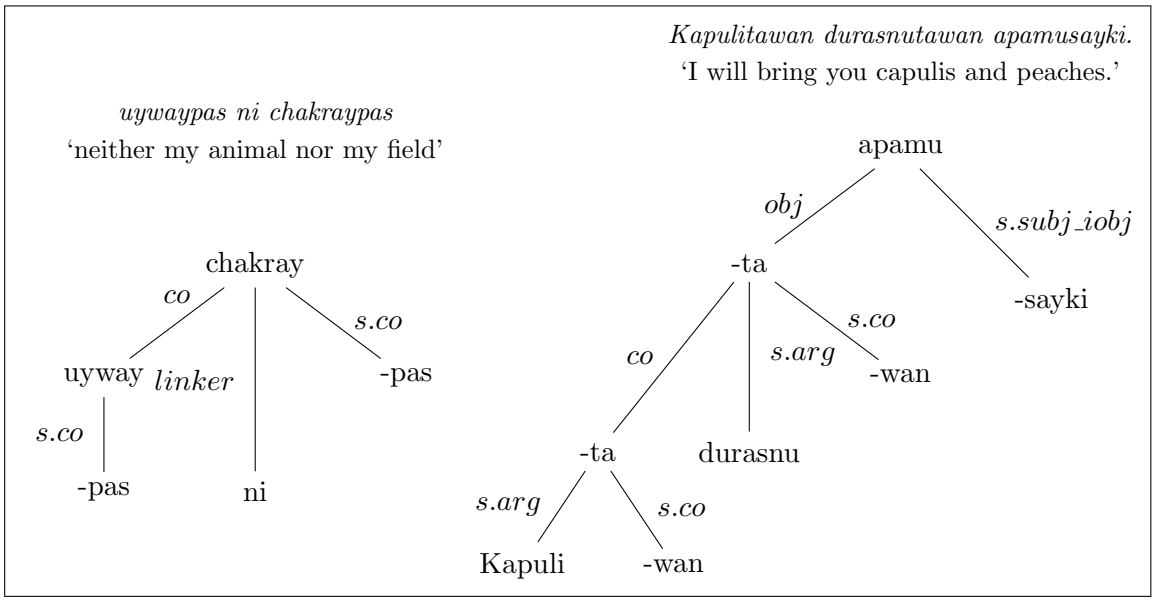


FIGURE 3.2: Annotation of Coordination

(20) *Kapuli -ta -wan durasnu -ta -wan apa -mu -sqayki.*
Capuli -ACC -CON peach -ACC -CON carry -DIR -1.SG>2.SG.FUT
'I will bring you capulis and peaches.'

[Cusihuamán 1976:142]

A further strategy for coordination in Quechua is the juxtaposition of two unmarked elements, e.g. *tayta mama* - 'parents' (lit. father mother) or *tuta p'unchaw* - 'night and day'.[17]

Modifiers of coordinated elements that modify all elements in the coordination are annotated as *secondary edges*. The modifier depends on its closest head but is attached via secondary edge to all the other elements it modifies, see the annotation of examples (21) and (22) in Fig. 3.3.

(21) kay *-pi llamk'a -nku, puklla -nku ima*
that -LOC work -3.PL play -3.PL too
'they work and play here (*kaypi* = *here*)

[17]Note that this construction can also be viewed as a nominal compound. For reasons of simplicity however, we annotate only roots that are written together as compounds (e.g. *pachamama* - 'Mother Earth', consisting of *pacha* - 'earth' and *mama* - 'mother'). For a more detailed analysis of nominal structures in Quechua see Floyd [2011].

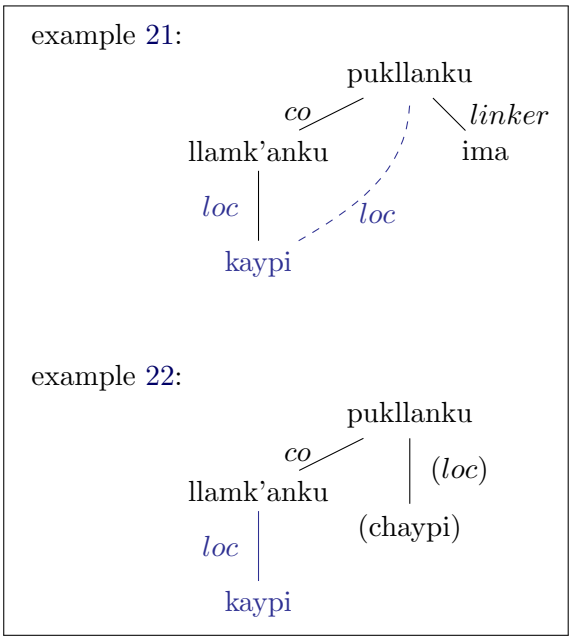


FIGURE 3.3: Annotation of Modifiers in Coordinations

(22) kay *-pi* *llamk'a -nku,* *(chay -pi)* *puklla -nku*
that -LOC work -3.PL this -LOC play -3.PL
'they work here and play [there]'

Another important issue in the treatment of coordinations is the possible elision of the finite verb: if the elided verb does not share all modifiers and arguments of the verb, we have to assume two individual coordinated clauses, consider examples (23) and (24). In the second clause of example (23) the verb *munan* is elided. A coordination on the level of the modifiers *p'unchaykunalla* and *p'unchaykunachu* is not possible due to the negation in the second clause that depends on the (elided) verb. In this case, we insert a dummy element to represent the missing verbal root *MUNA*. All the arguments and modifiers that this dummy shares with the first clause are linked to the dummy through secondary edges (dotted), see the annotation in Fig. 3.4.[18]

[18]Figure 3.4 illustrates the dependency tree as it is displayed in our annotation tool: black = word or morpheme, red = edge label, blue = morphology tag, purple = morphological class. For the meaning of tags and labels, see the list of abbreviations on page xiii.

(23) *Khayna Pachamama [kan] kay p'unchaykunalla muhuta munan, mana qollori p'unchaykunachu.*

Khayna Pachamama kay p'unchay -kuna -lla muhu -ta muna -n, mana
like.this mother.earth that day -PL -LIM seed -ACC want -3.SG not
qollori p'unchay -kuna -chu.
qollori day -PL -NEG

'Mother earth [is] like that, she wants the seed only on those days, not on the others that are *qollori*'. [Valderrama Fernandez and Escalante Gutierrez 1977]

(24) *Iskay laymi kaq [kan] sapanka laymipitaq papa tarpukuq (kan) achkha ladopi mana hukllapichu.*

Iskay laymi ka -q [kan], sapa -nka laymi -pi -taq papa tarpu -ku
two laymi be -AG [is] each -DISTR laymi -LOC -CON potato sow -RFLX
-q [kan] achkha lado -pi mana huk -lla -pi -chu.
-AG [is] much site -LOC not other -LIM -LOC -NEG

'[There used to be] two *laymis*, in each of them the potatoes used to be sowed in different places, never in only one place.'

[Valderrama Fernandez and Escalante Gutierrez 1977]

In example (24) the elided verb in the second clause is *tarpu-*, but as this is a habitual past form, we have to insert another dummy for the copula *kan*, see Fig. 3.5.

3.3.4 Focus

The evidential suffixes *-mi, -si* and *-cha* are usually attached to the focalized element, and thus besides their evidential function also serve as discourse markers. In yes/no-questions and negation, the interrogative/negation suffix *-chu* is attached to the focalized element [Sánchez 2010:47].

In their focalizing function, the evidential suffixes and *-chu* contrast with the topic markers *-qa* and *-ri*, which occupy the same slot in the suffix sequence and are mutually exclusive with the evidentials. Consider the following examples:

Evidential as focus marker:

(25) *Pawlucha -m wayqi -y -qa.*
Pablito -DIRE/FOC brother -1.SG.POSS -TOP
'My brother is Pablito.'

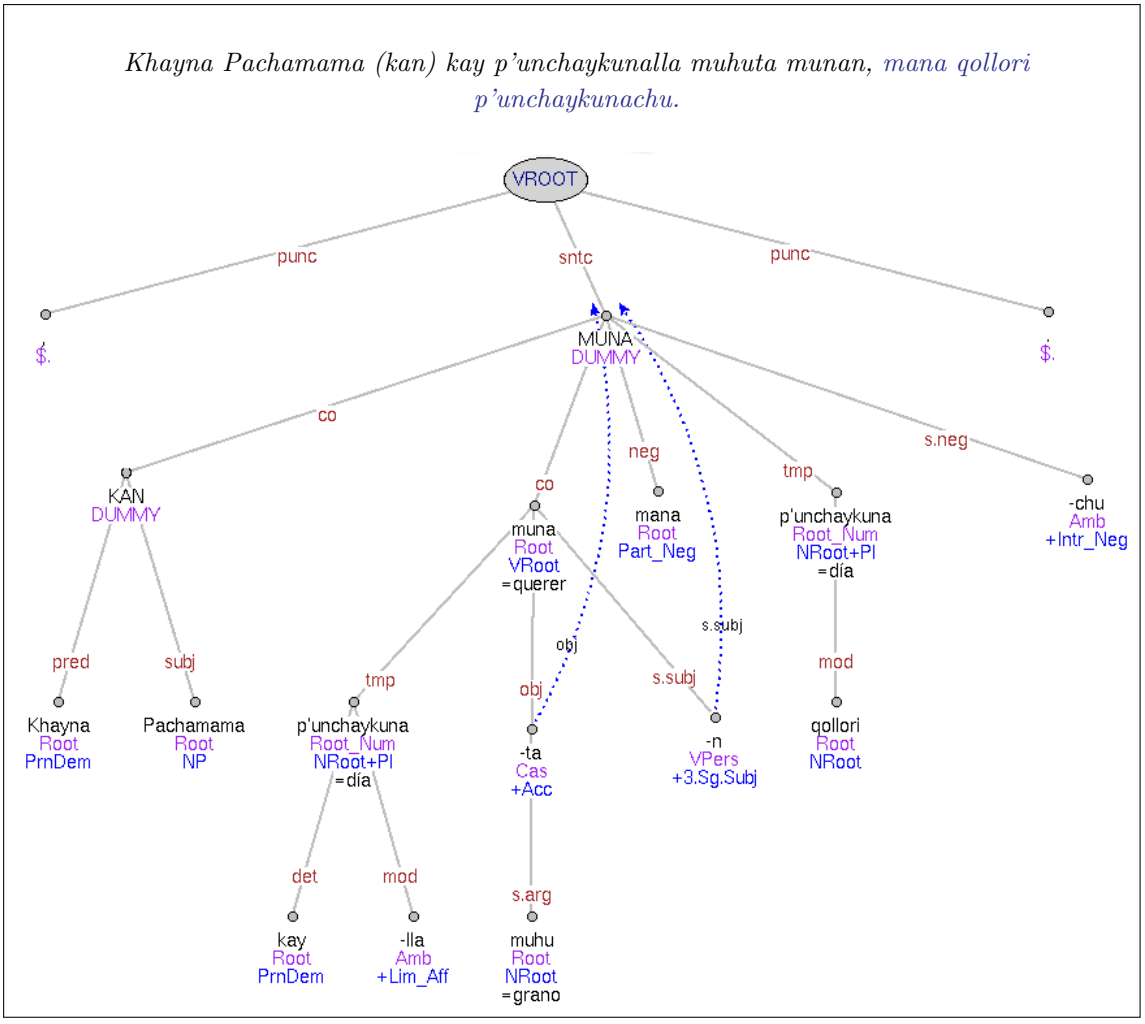


FIGURE 3.4: Finite Verb Elision in Coordination

(26) *Pawlucha -qa wayqi -y -mi.*
Pablito -TOP brother -1.SG.POSS -DIRE/FOC
'As for Pablito, he's my brother.'

Negation suffix as focus marker:

(27) *Mana -m huwis -chu ñuqa -qa ka -ni.*
Not -DIRE judge -NEG/FOC I -TOP be -1.SG
'I am not a judge (my profession is something else).'[19]

(28) *Mana -m ñuqa -chu huwis -qa ka -ni.*
Not -DIRE I -NEG/FOC judge -TOP be -1.SG
'The judge is not me (the judge is someone else).'

[19] *huwis*: from Spanish *juez* - 'judge'

Iskay laymi kaq (kan), sapanka laymipitaq papa tarpukuq (kan) askha ladopi mana hukllapichu.

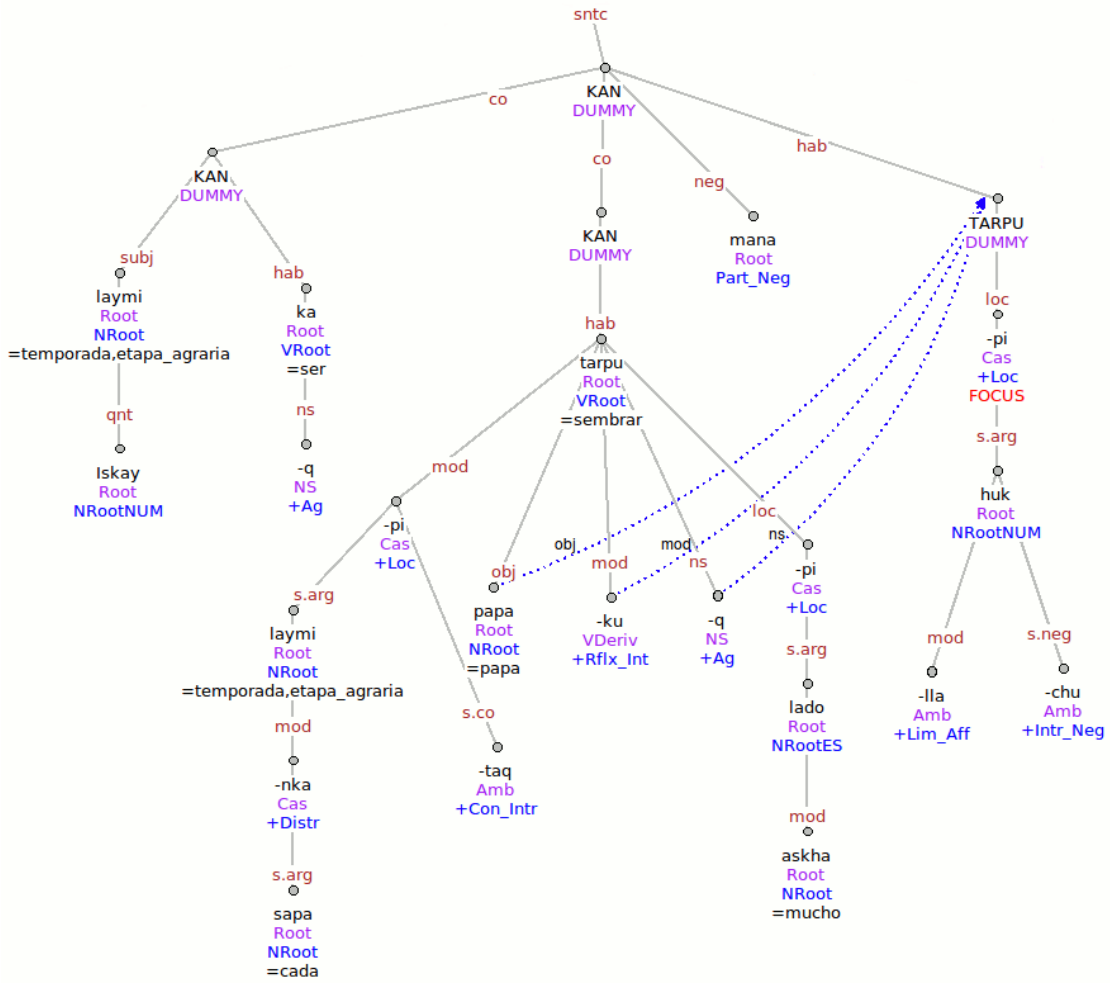


FIGURE 3.5: Finite Verb Elision in Coordination in Habitual Past

[Cusihuamán 1976:93]

This morphological syncretism of two functions in a single morpheme has to be represented accordingly in the dependency tree: as evidentials, they modify the clause as a whole, and therefore should depend on the head of the clause.[20] Nevertheless, as focus markers, the evidentials clearly belong to the element they are attached to. In order to represent both functions, we introduce an additional attribute *discourse* to the terminal nodes, which takes the value *FOCUS* if the element bears an evidential, or one of the

[20]The occurrence of evidentials is restricted to one per clause, as there cannot be more than one data source for an utterance.

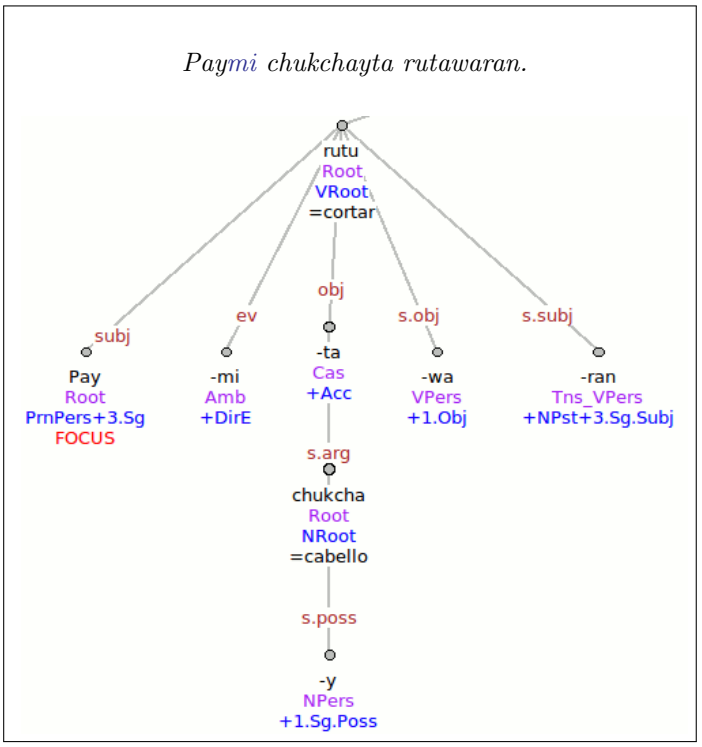


FIGURE 3.6: Annotation of Focus and Evidentiality

other focus markers. The evidential itself depends on the head of the clause, see also the annotation of example (29) in Fig. 3.6.

(29) *Pay -mi chukcha -y -ta ruta -wa -rqa -n.*
She -DIRE hair -1.SG.POSS -ACC cut -1.OBJ -PST -3.SG
'She cut my hair.'

[Valderrama Fernandez and Escalante Gutierrez 1977]

The situation with the interrogative function of *-chu* is similar: as focus marker, it relates to the element it is attached to, but as interrogative suffix, it modifies the clause as a whole. As with the evidentials, we set the value for the attribute *discourse* of the focalized element to *FOCUS*, while annotating *-chu* as direct dependent to the head of the clause.

3.3.5 Relative Clauses

Relative clauses in Quechua are nominal forms that are either agentive (nominalizer *-q*) or non-agentive (nominalizer *-sqa* or *-na*), for a more detailed description of these forms

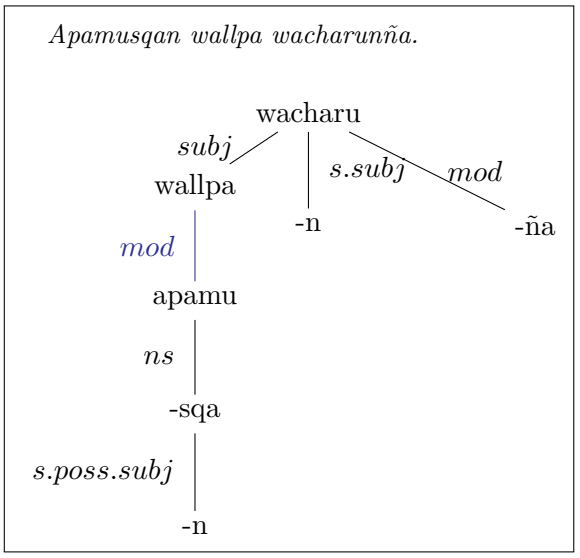


FIGURE 3.7: Relative Clause with External Head (*wallpa*)

see section 5.3.1. Canonical relative clauses precede the noun they modify, see examples (30) - (32), and are annotated with the label *mod*, see Fig. 3.7 with the annotation of example (31).

(30) *Aynikuq runakuna rumikunata pallanku chakramanta [..]*

Ayni -ku -q runa -kuna rumi -kuna -ta palla -nku chakra -manta..
help -RFLX -AG people -PL stone -PL -ACC pick.up -3.PL field -ABL

'The people who help pick up the stones from the field [..]'

[Soto Ruiz 2006:408]

(31) *Apamusqan wallpa wacharunña.*

Apa -mu -sqa -n wallpa wacha -rqu -n -ña.
bring -DIR -PERF -3.SG.POSS chicken give.birth -RPTN -3.SG -DISC

'The chicken he/she brought already laid eggs.'

[Soto Ruiz 1976:137]

(32) *¿Maypitaq llamk'anayki wasi kanqa?*

May -pi -taq llamk'a -na -yki wasi ka -nqa?
where -LOC -CON work -OBL -2.SG.POSS house be -3.SG.FUT

'Where is the house in which you will work?'

[Soto Ruiz 1976:137]

3.3.6 Internally Headed Relative Clauses

Internally headed relative clauses are a subtype of relative clauses and are defined as follows:

> An internally headed relative clause (IHR) is a subordinate clause which semantically modifies one of its own constituent nominals.
>
> [Hastings 2004:32]

In a dependency annotation, this leads to a dilemma: the head of the relative clause is also an element of the relative clause itself. Examples (33) and (34) illustrate the difference between internally headed (IHR) and externally headed (EHR) relative clauses:

(33) EHR: *Juanpa rantisqan wakaqa yuraqmi karqan.*

Juan -pa ranti -sqa -n ***waka*** *-qa yuraq -mi ka -rqa -n.*
Juan -GEN buy -PERF -3.SG.POSS cow -TOP white -DIRE be -PST -3.SG

'The cow that Juan bought was white.'

(34) IHR: *Juanpa waka rantisqanqa yuraqmi karqan.*

Juan -pa ***waka*** *ranti -sqa -n -qa yuraq -mi ka -rqa -n.*
Juan -GEN cow buy -PERF -3.SG.POSS -TOP white -DIRE be -PST -3.SG

'The cow that Juan bought was white.'

[Hastings 2004:55]

The relative clause in example (33) has an external head and can be annotated as described in the previous paragraph: the relative clause *Juanpa rantisqan* depends as *mod* on the head *waka*. However, the annotation of example (34) is not as straightforward: the head *waka* is embedded in the relative clause as it is clearly the object of the verb *ranti-* 'to buy', but *waka* is also the head of the relative clause and the subject of the main verb *ka-*. However, a dependency relation is defined as binary, therefore a node cannot have two heads. In order to annotate internally headed relative clauses, we insert a dummy node that represents the external head with the attribute *ExtHead:label* where *label* points to the head within the relative clause (*obj* in case of *waka*). See the annotation of example (34) in Fig. 3.8.

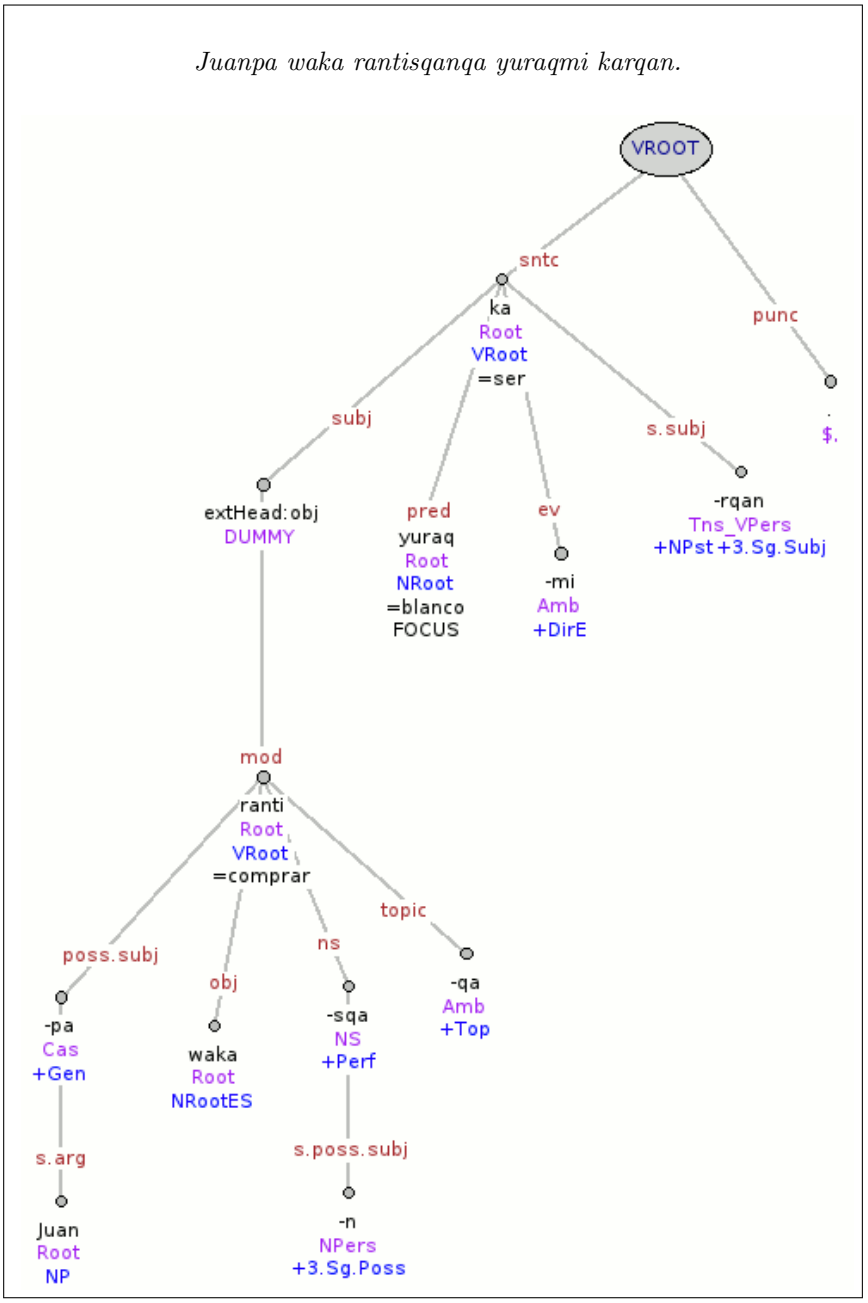


FIGURE 3.8: Relative Clause with Internal Head (*waka*)

3.3.7 Embedded Clauses

Subordinated finite clauses can be embedded with a demonstrative pronoun that resumes the content of the clause and bears the case suffix that links the clause to the main clause. Consider the following examples:

(35) *Tapuylla tapurikuy [maymantacha kani] chayta.*

Tapu -y -lla tapu -ri -ku -y may -manta -cha ka -ni
ask -INF -LIM ask -INCH -RFLX -2.SG.IMP where -ABL -ASMP be -1.SG
chay -ta.
this -ACC

'Could you please investigate as to where I am from?'

(Spanish: *Si fuera posible, averigua no más acerca de cuál es el lugar de mi procedencia.*)

(36) *Chaypi wachachaqa allinta yupaykun [llapanchus uwihakuna hampurqan*[21] *icha mayqinchus qhipakamurqanpis] chayta.*

Chay -pi wacha -cha -qa allin -ta yupa -yku -n llapan -chu -s
this -LOC girl -DIM -TOP good -ACC count -AFF -3.SG all -INTR -INDE
uwiha -kuna hampu -rqa -n icha mayqin -chu -s qhipa
sheep -PL come.back -PST -3.SG or how.many -INTR -INDE stay.back
-ka -mu -rqa -n -pis chay -ta
-RFLX -DIR -PST -3.SG -ADD this -ACC

'There the little girl begins to count well [if] all the sheep came or [if] some remained behind.'

(Spanish: *Y allí la chiquilla recuenta atentamente, para saber si todas las ovejas han venido o es que se han quedado algunas (en el campo).*)

[Cusihuamán 1976:266-267]

In this case, the finite clause depends on the resumptive pronoun as *rep* (repeated element) while the case suffix (or pronoun, if nominative) depends on the main verb labeled according to its function, see Fig. 3.9 with the annotation of example (35).

[21] About the analysis of *hampu-* - 'to come back': a complete analysis would be *ha -m -pu*, where *-m* is the short form of the directional *-mu* and *-pu* is the regressive ('back'). However, since there is no root **ha-* in modern Cuzco Quechua, we treat *hampu-* as a lexicalized unit.

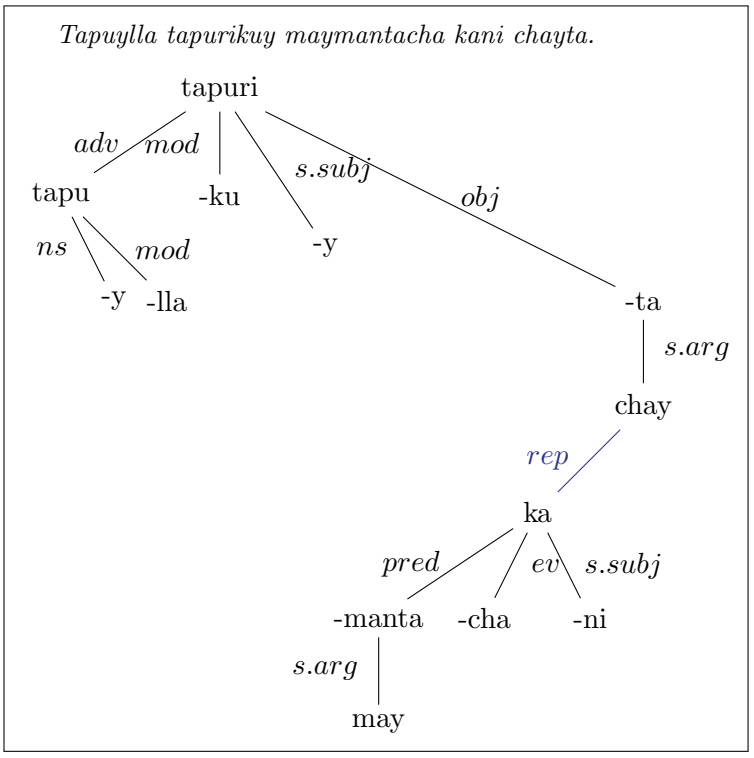


FIGURE 3.9: Embedded Clause with Resumptive Pronoun

3.4 Annotation Process

For the syntactic annotation we used the tool TrEd, a graphical tree editor that is not limited to a specific annotation scheme, but instead can display any tree-like structure in the XML-based format PML (Prague Markup Language).[22] The annotation scheme is defined in a PML schema, in the same way a DTD describes XML data, while the appearance of the trees is determined in a style sheet [Bejček et al. 2010:54]. The exact structure of the trees and how they are displayed is thus highly customizable. Furthermore, TrEd makes it possible to adapt pre-defined functions for handling trees in so called macros, so that not only the structure of the annotation but also the user menu and available functions for editing can be freely adjusted to the requirements of a given annotation scheme.

As the Quechua treebank is built on morpheme groups instead of words, the first step towards annotated trees is splitting the word forms into the required units with the

[22] https://ufal.mff.cuni.cz/tred/

finite-state processing described in section 2.3. Words that have more than one possible segmentation are disambiguated with the conditional random fields approach, see section 2.4 for details.

In the next step, we convert the resulting text to PML, the XML-based format required by the graphical annotation tool TrEd. In order to speed up and facilitate the manual annotation, a rule-based script pre-annotated some of the simple structures that can easily be covered by rules, such as the case suffixes and their nominal arguments. As all these pre-processing steps were done fully automatically, there will inevitably be wrong structures that have to be corrected during the manual annotation process. Nevertheless, the automatic pre-processing works well enough to facilitate the manual annotation.[23]

Once the manual annotation is finished, a second annotator checks the trees and makes corrections if necessary. Furthermore, we use the PML tree query [Štěpánek and Pajas 2010] to check the consistency of notoriously difficult structures, such as the internally headed relative clauses or the embedded clauses described in section 3.3. Figure 3.10 shows an overview of the annotation process for our Quechua treebank.[24]

So far, this chapter has dealt with the manual annotation of Quechua dependency trees. The following section will take a step further and illustrate how we can speed up the annotation process by annotating Quechua sentences automatically with dependency trees.

3.5 Parsing Quechua Sentences

Parsing in a linguistic context refers to the task of automatically analyzing the syntactic structure of a sentence according to a specific grammar formalism. In the context of a dependency grammar, such as the annotation scheme for Quechua presented in this

[23] Annotation time varies depending on the complexity of the sentence, but generally our annotator in Peru completed 4-10 sentences in one hour.

[24] The treebanks are accessible through the PML tree query web interface at `http://marvin.ifi.uzh.ch/kitt/pmltq/`. The treebank, including the TrEd macro, PML schema and style sheet are freely available from the project's download site at `http://kitt.ifi.uzh.ch/kitt/squoia/download.html` or as part of the multilingual treebank SMULTRON (`http://kitt.ifi.uzh.ch/kitt/smultron/`). These packages are not updated regularly, but the actual versions are available from the project code repository at `https://github.com/ariosquoia/squoia/tree/master/`.

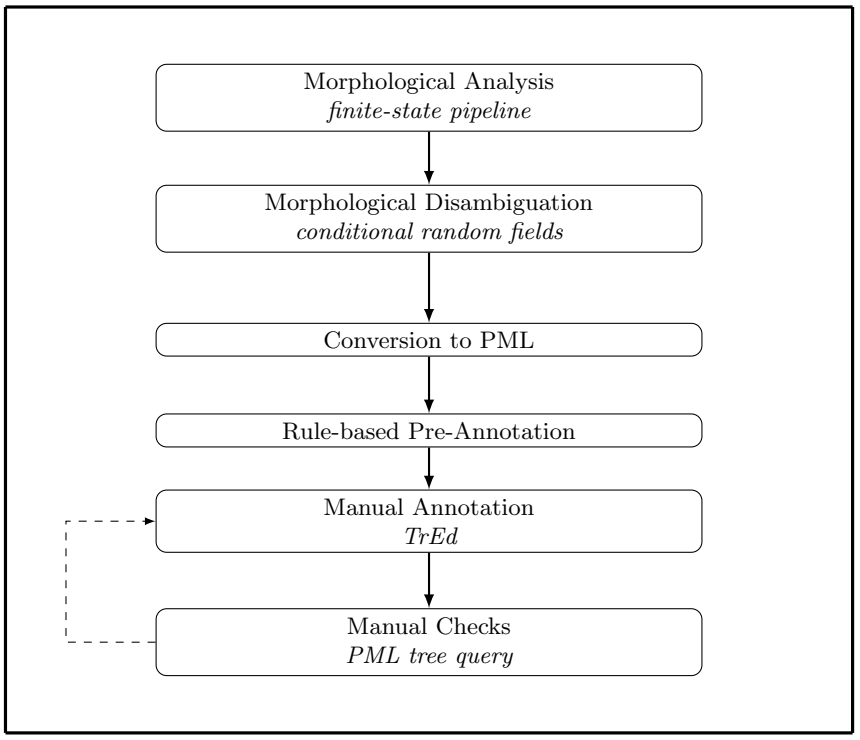


FIGURE 3.10: Annotation Process

chapter, parsing means mapping an input sentence S to its dependency graph G [Kübler et al. 2009:6].

This syntactic analysis can be based on grammatical rules or on a statistic model, or a combination of both. A rule-based parser requires a set of hand-written rules that describe the grammatical structures of the language according to the chosen syntax formalism. Writing such a rule-based grammar that covers most of the grammatical sentences in a given language is a time-consuming task that can only be implemented at great expense. On the other hand, training a statistical parser on annotated syntax trees can be done with a relatively small effort if such data is available.

3.5.1 Conversion PML to CoNLL

The Quechua treebank created in the SQUOIA project contains about 2000 sentences with full dependency annotations and can thus serve as training material for a statistical parser. However, annotations that involve artificial dummy elements need to be considered carefully and transformed into a representation without artificial elements. A common format for dependency annotation is the tabbed CoNLL format used in the

shared tasks of the Conference on Natural Language Learning[25], and since most available statistical parsers accept CoNLL as input, the first step towards a Quechua parser is the conversion of the PML trees to the CoNLL format.

There are different cases where our annotation scheme relies on artificial dummy elements in order to create a dependency tree:

1. sentences where the 3rd person singular form of the copula *ka-* is omitted, those include:

   a. predicative clauses:

      Allin runa -m chay wiraqucha -qa [kan].
      good man -DIRE this mister -TOP [is]

      'This man [is] a good person.'

      [Cusihuamán 1976:137]

   b. habitual past 3rd person clauses:

      Nisyu -ta maqa -wa -q [kan].
      too.much -ACC beat -1.OBJ -AG [is]

      'He used to beat me too much.'

      [Valderrama Fernandez and Escalante Gutierrez 1977]

   c. obligative present clauses:

      Achkha tajamal -kuna -ta -qa ruwa -na -y [kan].
      much tamal -PL -ACC -TOP make -OBL -1.SG.POSS [is]

      'I have to make many tamales.'[26]

      [Valderrama Fernandez and Escalante Gutierrez 1977]

2. verb ellipsis in coordinations, see examples (23) and (24) on page 50
3. internally headed relative clauses, see example (34) on page 56

In all cases listed under (1), we can drop the dummy and use another element as head: in predicative clauses, such as (a), this will be the predicative element (*runa*), while in habitual past and obligative present forms, such as (b) and (c), the nominalized main verb can take the place as head of the sentence.

[25] see `http://ifarm.nl/signll/conll/`
[26] A corn-based meal.

On the other hand, verb elision in coordinations (type 2) and internally headed relative clauses (type 3) cannot be displayed as proper dependency trees without artificial elements. Basically, these necessary dummy elements can be handled in parsing in three ways [Seeker et al. 2012]:

- insert dummies before parsing
- create special labels for the dependents of dummy elements that contain information about the missing head
- allow the parser to insert dummy elements

In statistical parsing, all of these options require sufficient data for learning where to insert empty elements or special labels. Since the Quechua treebank built in SQUOIA contains only ∼2000 sentences, the number of clauses with dummies of type 2 and 3 is far too small for any data-driven approach: the whole treebank contains only 11 coordinations with elided heads and 21 internally headed relative clauses.

For this reason, only sentences with a dummy element of type 1 were converted to CoNLL, whereas the few clauses with dummy elements of type 2 or 3 were discarded for the parsing experiments.

3.5.2 Parsing and Preliminary Evaluation

We chose to use the MaltParser [Nivre et al. 2007], since it is a freely available, language-independent system that comes with a module for (semi-) automatic optimization, the MaltOptimizer [Ballesteros and Nivre 2015]. As the MaltParser system includes nine different parsing algorithms, each with several possible features to be set, which furthermore can be combined with different machine learning algorithms, tuning the MaltParser manually is a complex task that requires in-depth knowledge about the system. The MaltOptimizer offers an easy way to find suitable settings for their specific data.

Ideally, optimization should be done on a held-out development set. However, our Quechua treebank in CoNLL format contains less than 2000 sentences, a number far too small to create a development set of reasonable size. As a consequence, optimization was done on the whole data set with cross-validation instead of a development set. For this

	LAS	UAS	Label accuracy
standard settings	70.52	79.04	73.33
MaltOptimizer settings	82.17	88.46	86.70

TABLE 3.1: Preliminary Results with MaltParser (10-fold Cross-Validation)

reason, the results presented here have to be considered with caution, the performance of the parser on completely new data might not be as good.

Table 3.1 contains the results obtained with the standard settings as opposed to the optimized settings suggested by MaltOptimizer. For both settings, labeled attachment score (LAS), unlabeled attachment score (UAS) and label accuracy were calculated through 10-fold cross-validation.

A complete package for morphological analysis and parsing is available from our website.[27]

Table 3.2 illustrates the performance of the parser on different dependency relations: certain tokens, such as nominalizing suffixes and subject markers, are almost unambiguous and thus easy to handle for the parser. On the other hand, the most difficult labels for the parser turn out to be connected to ambiguous tokens that are difficult for human annotators as well. For instance, the instrumental case suffix *-wan* - 'with' can imply company (label=acmp), an instrument (label=instr) or coordination (label=s.co). The last case is relatively clear since every coordinated element will bear the suffix *-wan* but the other two uses of *-wan* are harder to distinguish:

(37) acmp: *Huq wiraquchakuna**wan**mi llamk'ani.*

*Huq wiraqucha -kuna **-wan** -mi llamk'a -ni.*
Other mister -PL -INSTR -DIRE work -1.SG

'I work with some other men.' [Cusihuamán 1976:126-127]

[27] https://github.com/ariosquoia/squoia/releases

Label	Meaning	Recall	Precision	F-Measure	Frequency (in treebank)
highest:					
ns	nominalizer	99.93	99.91	99.92	5860
topic	topic suffix	99.8	99.6	99.70	1579
s.subj	subject suffix	99.81	99.42	99.61	2128
sentence	root node	100	99.19	99.60	1923
s.poss.subj	possessive suffix on nominalized noun (subject)	98.79	97.82	98.30	1384
s.poss	possessive suffix	97.01	98.34	97.67	1057
ev	evidentiality	97.75	97.36	97.55	1503
s.arg	argument of a (case) suffix	96.2	97.11	96.65	4512
s.arg.claus	nominalized clause argument of a case suffix	97.18	94.29	95.71	2221
det	attributive demonstrative pronoun	93.5	94.77	94.13	431
neg	negation	94.8	92.92	93.85	355
s.obj	object suffix	96.61	88.37	92.31	126
s.neg	negation suffix	98.01	86.03	91.63	215
poss.subj	possessor subject (nominalized clause)	92.23	90.43	91.32	544
s.co	coordinative suffix	92.24	88.92	90.55	1395
flm	foreign language material	89.21	86.68	87.93	547
lowest:					
rep	repeated element	36.92	43.64	40.00	82
acmp	company ('with')	40.74	37.08	38.82	71
hab	habitual past	29.66	51.19	37.56	156
dm	discourse marker	35.29	34.29	34.78	35
s.subj_iobj	portmanteau suffix for subject and object person	20.0	100.0	33.34	24
oblg	obligation	19.53	50.0	28.09	117
dupl	reduplicated (truncated) root	17.95	63.64	28.00	35
voc	vocative	18.18	28.57	22.22	18
numord	ordinal number	10.0	33.33	15.38	8
app	apposition	9.3	30.77	14.28	86
instr	instrument ('with')	9.76	16.0	12.12	33
arg	oblique argument of verb	7.69	20.0	11.11	17
par	parenthesis	4.69	16.67	7.32	76
distr	distributive	0	0	–	5
r.disl	right dislocation	0	0	–	6

Table 3.2: F-Measure Dependency Relations

(38) instr: *Ñawinchik**wan**mi qhawanchik, siminchik**wan**taq rimanchik.*

Ñawi -nchik ***-wan*** *-mi* *qhawa -nchik,* *simi* *-nchik*
eye -1.PL.INCL.POSS -INSTR -DIRE see -1.PL.INCL mouth -1.PL.INCL.POSS
-wan *-taq* *rima* *-nchik*
-INSTR -CON speak -1.PL.INCL

'We see with our eye(s) and we speak with our mouth.' [Soto Ruiz 1976:82-83]

Furthermore, some of lowest ranking labels in Table 3.2 have a very low frequency: for instance, ordinal numbers are expressed in Quechua by the number followed by either *kaq* or *ñiqin*. Ordinal numbers occur 6 times with *kaq* and only twice with *ñiqin* in the treebank. While *ñiqin* has no other function than marking ordinal numbers, *kaq* is the agentive form of the copula and as such occurs in many other functions throughout the treebank. Ordinal numbers are simply too sparse in our data for the parser to learn. The same holds true for right dislocations (r.disl) and the distributive usage of the case suffixes *-kama* and *-nka*. On the other hand, the distinction between parenthesis (par) and appositions (app), same as with *-wan*, is often difficult as well for human annotators. Basically, we consider an apposition to be an explanation or a more detailed description of a given entity. As a rule of thumb, appositions can replace the entity they describe without affecting the grammaticality of the sentence. Consider example (39): the whole clause to the right is an apposition to the phrase *chay historiapi* - 'about this story', and since the head *munasqanpi* bears the same case suffix, it can replace *chay historiapi* and the resulting sentence is still grammatical.

A parenthesis, on the other hand, is an insertion of additional information that cannot be part of the syntactic structure of the clause. Consider the phrase *estanciapiqa pisipunin mikhuna* - 'food was always scarce at the farm' in example (40): this insertion is additional information, a commentary by the speaker that cannot otherwise be included into the syntactic structure of the sentence.

(39) apposition: *Chay historia**pi**m ñuqa pensarqani: Inka Qusquta ruwachkaspa p'unchawta hatunyayachiy munasqan**pi**, Inka Qullap wintunmanta cuidakuspa.*

Chay historia -pi -m ñuqa pensa -rqa -ni: Inka Qusqu -ta ruwa
this story -LOC -DIRE I think -PST -1.SG Inka Cuzco -ACC make
-chka -spa p'unchaw -ta hatun -ya -yka -chi -y muna -sqa
-PROG -SS day -ACC big -AUTOTRS -AFF -CAUS -INF want -PERF
-n -pi, Inka Qulla -p wintu -n -manta cuida -ku -spa
-3.SG.POSS -LOC Inka Qulla -GEN wind -3.SG.POSS -ABL protect -RFLX -SS

'I thought about this story: The Inka, trying to make the day longer, building Cuzco, protecting [it] from Inka Qulla's wind.'

(40) parenthesis: *Wawankuna y paykuna imataq maltratawaqku, mikhunapas pisitaq, **estanciapiqa pisipunim mikhuna** mana p'unchayta tariqraqchu kani may-manpas ripunaypaq.*

Wawa -n -kuna paykuna ima -taq maltrata -wa -qku mikhuna
child -3.POSS -PL they too -CON mistreat -1.OBJ -3.PL.HAB food
-pas pisi -taq estancia -pi -qa pisi -puni -m mikhuna mana
-ADD little -CON farm -LOC -TOP little -DEF -DIRE food not
p'unchaw -ta tari -q -raq -chu ka -ni may -man -pas ripu -na
day -ACC find -AG -CONT -NEG be -1.SG where -DAT -ADD go.away -OBL
-y paq
-1.SG.POSS -BEN

'They and their children mistreated me, and food was scarce, food was always scarce at the farm, I couldn't find the day to go away to wherever [it would be].'

Valderrama Fernandez and Escalante Gutierrez [1977]

The performance of the parser on the different dependency relations thus correlates with data sparseness, but also with difficulties in human judgment for the more complex cases.

3.6 Summary

In this chapter, we introduced the framework of dependency grammar and we gave a detailed description of the specific annotation scheme that we developed for Quechua. For some features of Southern Quechua syntax, such as internally headed relative clauses or the morphological syncretism some of the independent suffixes exhibit in specific

contexts, there is no straightforward dependency annotation, since they encode relations that are not binary. In order to annotate these elements with their full information, we have to rely on artificial elements and additional attributes in the word nodes.

The final treebank contains 1979 syntactically and morphologically annotated sentences and was used to train MaltParser. The resulting parser performs reasonably well, given the sparse training data. For future annotations, the parser can provide a rough annotation that the user corrects, instead of annotating each sentence from scratch. This procedure will save time, since the parser handles frequent, repetitive structures reliably, the user can focus on the more difficult parts of the annotation. This chapter concludes the first part of the thesis that dealt with monolingual resources and tools for Quechua. Chapter II will explore the area of bilingual applications for Spanish and Quechua.

Part II

Bilingual Spanish-Quechua Resources

Chapter 4

Word-Aligned Parallel Text: Bilingwis Spanish-Quechua

4.1 Introduction

Parallel word-aligned text collections are an important resource not only for contrastive language studies, but also as support for translators, as they make it possible to search for words and their translations in different contexts. Furthermore, parallel documents are also useful for tasks such as word sense disambiguation, terminology extraction and cross-language corpus linguistics [Volk et al. 2011]. With corpora that are large enough, statistical word alignment is the preferred method to find the corresponding translations. The statistical alignment will never be error-free, but in turn it will also find rare translation options that might not be found in a conventional dictionary.

Bilingwis[1] is a system that allows to search for word translations in bilingual text, developed at the University of Zurich. The searchable corpora include the yearbooks of the Swiss Alpine Club (SAC) from 1957-2013 in German and French, but also Swiss law texts in German, French and Romansh. Furthermore, Bilingwis contains word-aligned speeches from the TED (Technology, Entertainment, Design) conference[2] in English and Chinese.

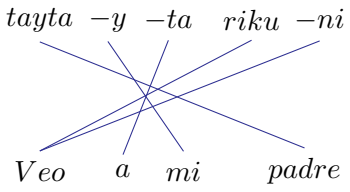
[1] see http://kitt.cl.uzh.ch/kitt/bilingwis
[2] 'TED talks', see http://www.ted.com/talks

4.2 Spanish-Quechua Bilingwis

In addition to the corpora mentioned above, Bilingwis contains the biography of Gregorio Condori Mamani and his wife in Spanish and Cuzco Quechua.[3]As the amount of parallel texts for the language pair Quechua-Spanish is not large enough for statistical alignment approaches, the word alignments have been annotated rule-based through the bilingual dictionary created for the translation system (see chapter 5.4). As opposed to the other parallel corpora in Bilingwis, the alignments in the Spanish-Quechua part are not based on words, but on morphemes, due to the fact that often, a Spanish word corresponds to a Quechua suffix, see the alignments between (41) and (41).

(41) *tayta -y -ta riku -ni.*
padre -1.SG.POSS -ACC riku -1.SG
'I see my father'

(42) *Veo a mi padre.*
See.-1.SG PREP my father
'I see my father.'



In order to create the alignments for Bilingwis, the first step is to align the sentences in the parallel texts. We used hunalign [Varga et al. 2005], a statistical tool to create the inital alignments. As sentence alignment is absolutely crucial for the quality of the resulting word alignments, we manually corrected these automatically calculated sentence pairs.[4] The next step was the morphological analysis and disambiguation of the Quechua texts (see Sections 2.3 and 2.4). The Spanish side of the corpus was morphologically analysed with FreeLing and tagged with Wapiti (see Section 5.2 for a detailed description). Both annotated texts were converted to XML, which serves as input for the rule-based alignment. Once the corresponding words and morphemes have been extracted from the text, they are loaded into the mySQL database that serves as back-end to Bilingwis.

Our rule-based word alignment uses a relatively fine-grained approach that excludes impossible alignments: inherently ambiguous Spanish words, such as prepositions, that

[3] http://kitt.ifi.uzh.ch/kitt/squoia/bilingwis_quz_es/
[4] We used InterText for this purpose, see Vondřička [2014].

correspond to either postpositions or suffixes in Quechua, are only aligned if their nominal argument is also aligned with the corresponding Quechua root. For instance, the Spanish preposition *a* can be translated in Quechua as accusative (*-ta*) or dative-illative suffix (*-man*), and there might be more than one possible alignment in a given sentence pair. For instance, see examples (43) and (44): the preposition *a* from the phrases ***a** la casa de mi madrina* and ***a** dónde* have 3 alignment candidates: *may**man***, *haykuy**ta*** and *wasin**ta***.

Spanish:

(43) *Así, ya en Acopia, no sabía **a** dónde entrar,*
Like.that already in Acopia not know.1.SG.PST **to** where enter
*tenía vergüenza de regresar **a** la casa de mi madrina.*
have.-1.SG.PST shame of come.back **to** the house of my godmother
'So, already in Acopia, I did not know where to go, I was ashamed of returning to my godmother's house.'

Quechua:

(44) *Ña -m Aqopiya -pi -ña mana may **-man** hayku -y **-ta***
Already -DIRE Acopia -LOC -already not where **-DAT/ILL** enter -INF **-ACC**
ati -ni -chu p'inqa -ku -y -niy -wan madrina
can -1.SG -NEG feel.shame -RFLX -INF -1.SG.POSS -INSTR godmother
*-y -pa wasi -n **-ta** mana kuti -yku -ni -chu.*
-1.SG.POSS -GEN house -3.SG.POSS **-ACC** not come.back -AFF
-1.SG -NEG

'Already in Acopia, I could not enter anywhere, I did not [want to] return to my godmother's house with shame.'

Instead of aligning all possible combinations, most of them wrong, we check whether the argument of the Spanish prepositions (*casa* and *dónde*) are aligned to one of the Quechua candidates' nominal roots (*may*, *hayku* and *wasi*). The algorithm will find that *may* is aligned to *dónde*, and create an alignment between the first *a* and the suffix *-man*, discarding all other candidates for this *a*. Likewise, *casa*, the nominal argument of the second *a*, is aligned to the Quechua root *wasi*, therefore we can be sure that this *a* has to be aligned to *-ta* in *wasinta*, see Fig. 4.1 for a graphical illustration of the aligned sentences.

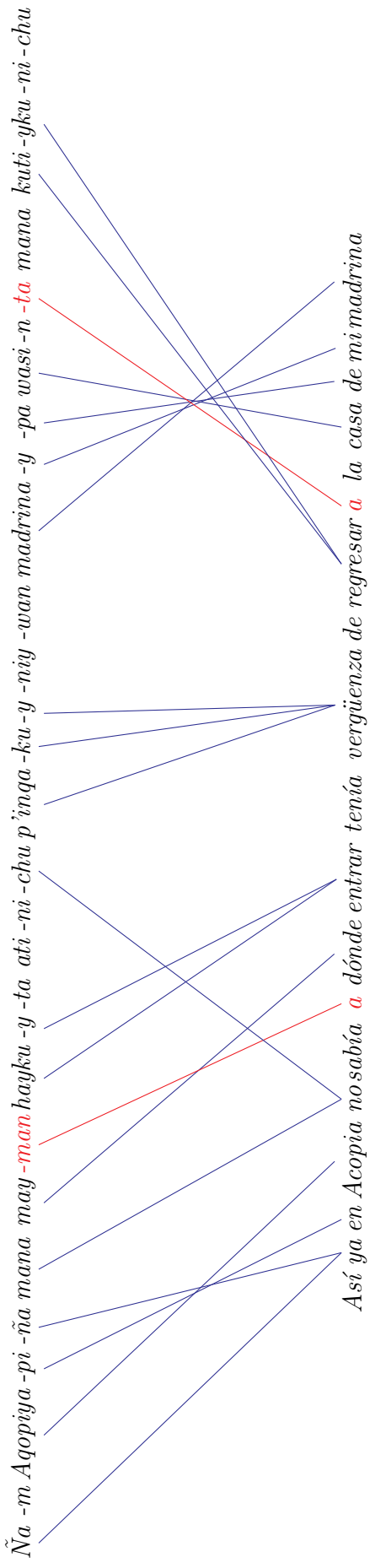


FIGURE 4.1: Alignments of Examples (43) and (44)

Figure 4.2 shows the results as they are displayed in Bilingwis for the search of the Spanish preposition *a*: in total, the Spanish preposition *a* corresponds 451 times to the accusative *-ta* and 345 times to the dative-illative *-man*.

Bilingwis can be searched in both directions and Figure 4.3 illustrates the results of the search for the Quechua root *yacha-* - 'to know': *yacha-* corresponds most frequently to Spanish *saber* - 'to know', but in combination with the causative *-chi*, the translation is usually *enseñar* - 'to teach'. Furthermore, there are some cases where *yacha-* corresponds to *aprender* - 'to learn' in Spanish.

In addition to the Quechua word forms, the data base behind Bilingwis contains the individual suffixes. This makes it possible to search not only for Quechua words, but also for single morphemes, see Fig. 4.4 with the results for the search of the Quechua suffix *-kama*. The case suffix *-kama* has two basic functions, it marks either an end point ('terminative'), see examples (45)-(46), or a feature that is shared by several individuals ('distributive'), see example (47). Furthermore, it occurs in subordinated clauses in combination with the nominalizing suffix *-na*, in this case, it indicates either an end point ('until') or simultaneity ('while'), see examples 48 and 49.

terminative:

(45) *Wayllay* ***-kama*** *-m* *ri* *-saq.*
Huayllay **-TERM** -DIRE go -1.SG.FUT
'I will go as far as Huayllay.'

(46) *Wata* ***-kama*** *-m* *mana hamu -nqaku* *-chu.*
year **-TERM** -DIRE not come -3.PL.FUT -NEG
'They won't come until [next] year.'

[Soto Ruiz 1976:80-81]

distributive:

(47) *Allin runa* ***-kama*** *-m* *ka -nku.*
good man **-DISTR** -DIRE be -3.PL
'They are all (each of them) good people.'

[Dedenbach-Salazar Sáenz et al. 2002:190]

'until':

(48) *Kay -lla -pi suya -chka -nki kuti -rqa -mu -na -y* ***-kama.***
that -LIM -LOC wait -PROG -2.SG return -RPTN -DIR -OBL -1.SG.POSS **-TERM**
'You wait here until I come back.'

[Cusihuamán 1976:210]

'while':

(49) *Sama -na -y* ***-kama*** *-m pay puklla -ku -chka -n.*
rest -OBL -1.SG.POSS **-TERM** -DIRE he/she play -RFLX -PROG -3.SG
'He/she plays while I rest.'

[Soto Ruiz 1976:81]

As the results in Fig. 4.4 show, the most frequent use of *-kama* is the marking of an end point, in this case, *-kama* corresponds to the Spanish preposition *hasta* - 'until'. In subordinated clauses, *-kama* denotes slightly more often simultaneity, in Spanish expressed through the conjunction *mientras*, than the end point of an action, in Spanish marked by the conjunction *hasta que* ('until').

4.3 Summary

In this chapter, we introduced the tool Bilingwis that allows to search for translations in a collection of parallel texts. Since statistical word alignment was not possible for the Spanish-Quechua language pair due to the small number of available parallel texts, we present a rule-based approach that relies on a dictionary for the alignment of words and morphemes. A further distinction from the other language pairs in Bilingwis is that we align not only whole words, but also morphemes or morpheme groups on the Quechua side. This allows for a better annotation of the correspondences.

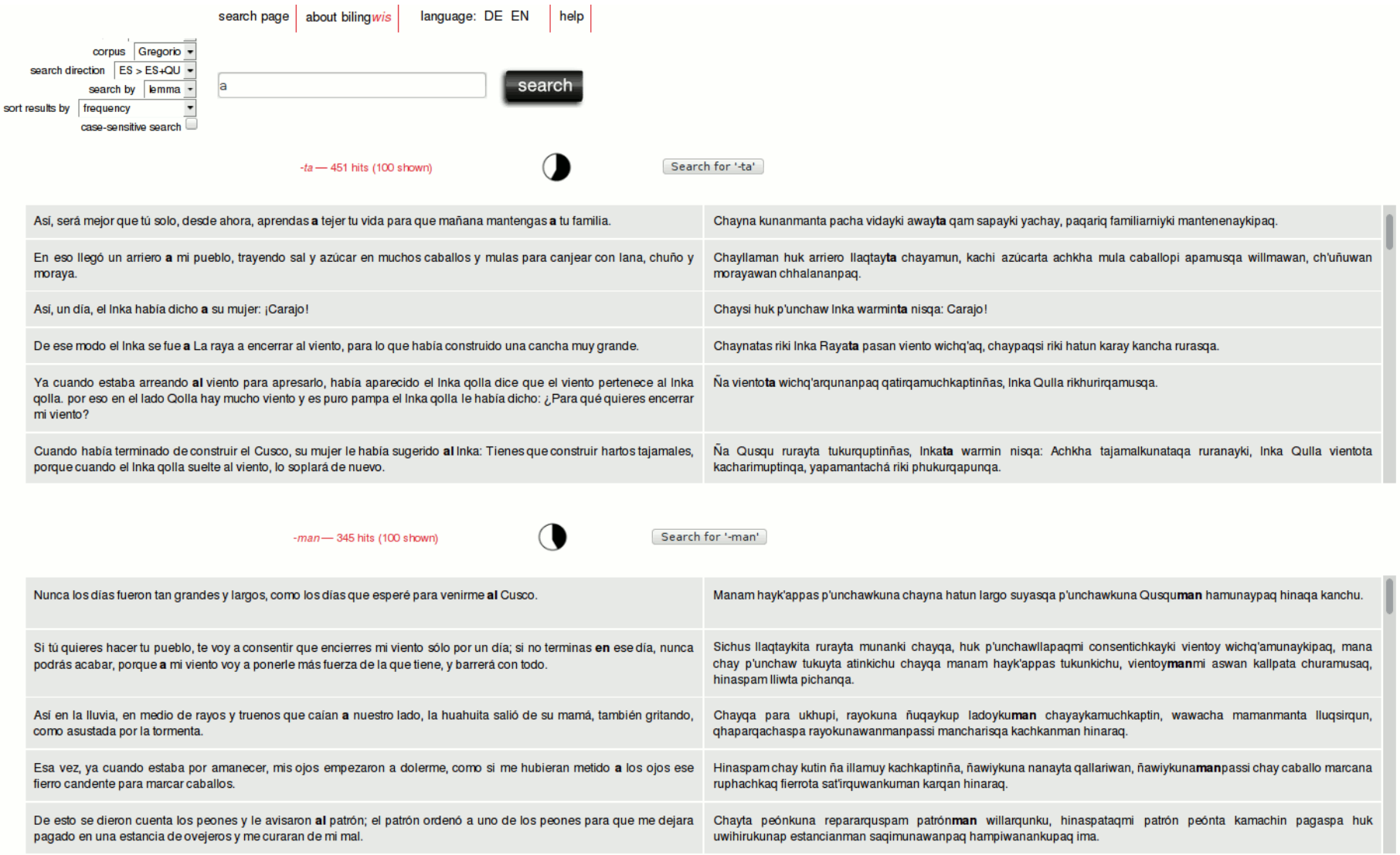


FIGURE 4.2: Bilingwis Results for the Spanish Preposition *a*

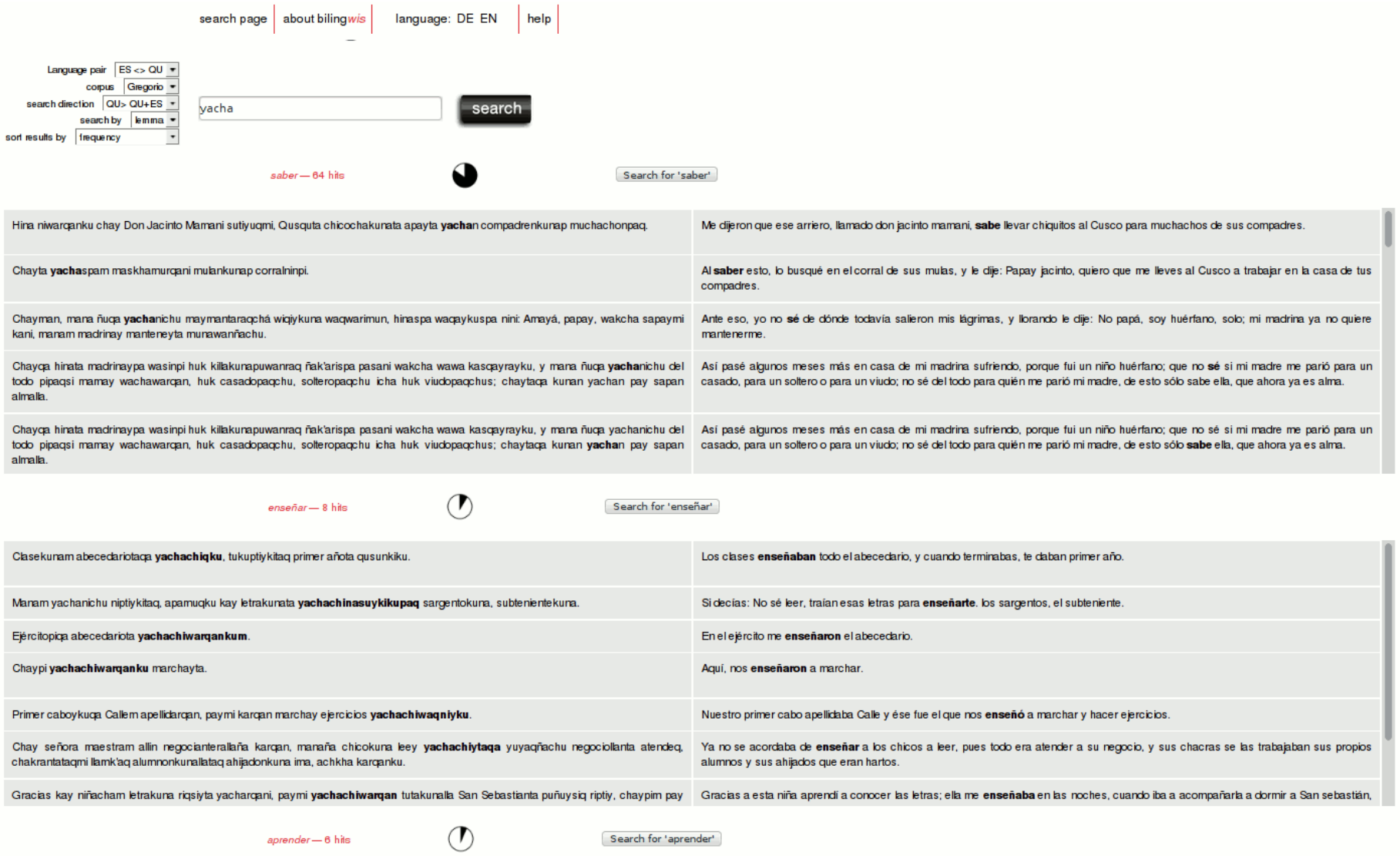


FIGURE 4.3: Bilingwis Results for the Quechua Root *yacha*

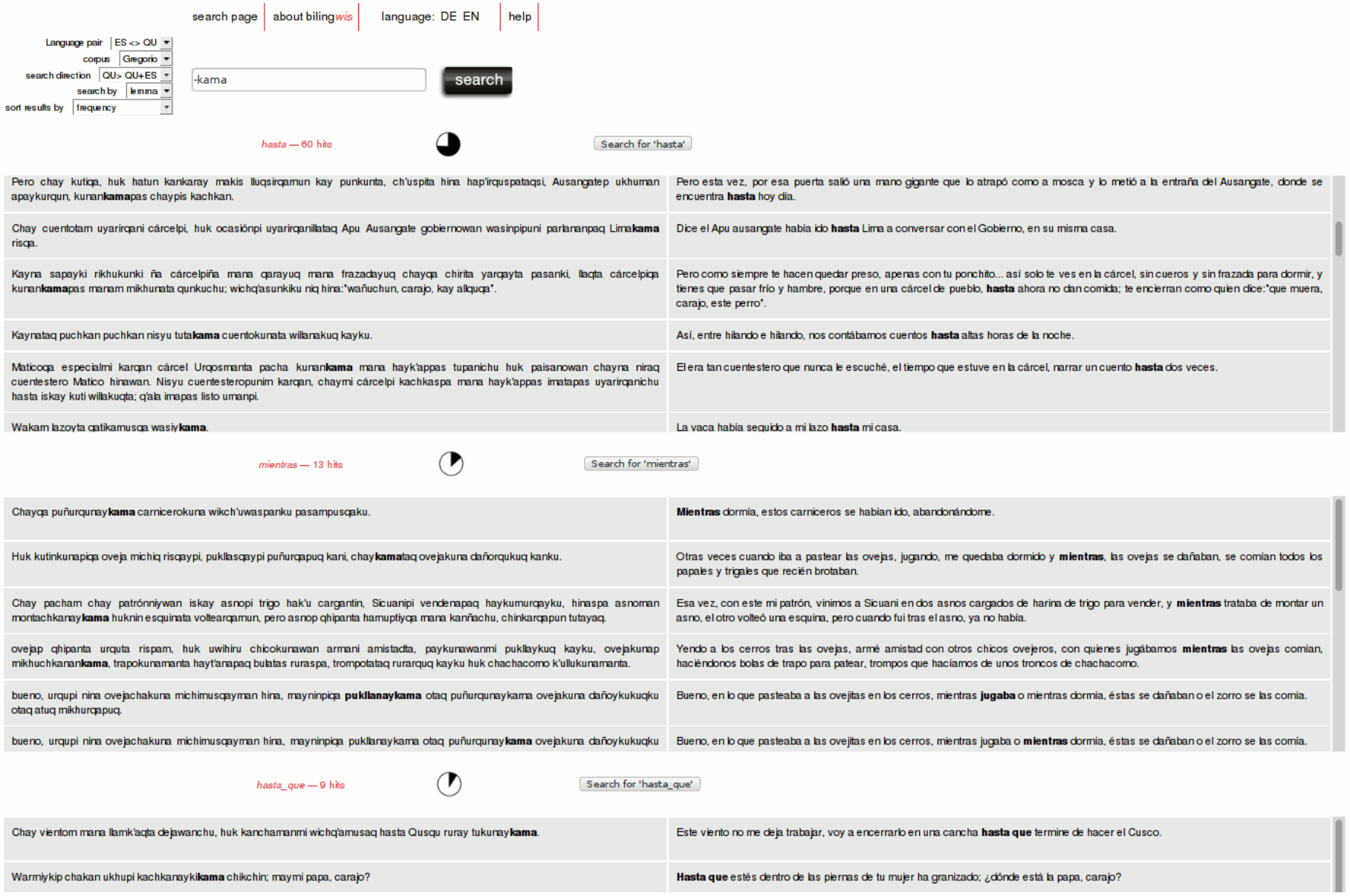


FIGURE 4.4: Bilingwis Results for the Quechua Suffix *-kama*

Chapter 5

Hybrid Machine Translation Spanish-Quechua

5.1 Introduction

Machine translation (MT) refers to the application of automatically translating text or speech from one language to another by computer. The first approaches in machine translation ranged from direct word by word translation over elaborated transfer methods with morphological and syntactic analysis, up to a more complex transfer through an abstract meaning representation referred to as interlingua [Koehn 2010:14-15]. In recent years, statistical machine translation has received much attention with the introduction of data-driven methods and the availability of large parallel corpora for many language pairs. Instead of relying on complex grammar and transfer rules, a statistical MT system learns how to translate from one language to another from a collection of human-translated texts. However, the performance of a statistical MT system depends heavily on the size and quality of the parallel corpus that it learns from, and for many language pairs, there is simply not enough human-translated material to build reliable statistics.

For the language pair Spanish-Quechua, the amount of parallel text is very small: since Quechua is still a predominantly spoken language, even monolingual texts are relatively scarce, limited mostly to books with traditional stories, legal documents (e.g. the Peruvian Constitution, Declaration of Human Rights) and translations of religious texts. For

this reason, a purely statistical approach to machine translation is not suitable for this language pair. Instead, the core of the translation system presented here is a classical, rule-based pipeline that relies on a deep analysis of the source sentence and a set of cascaded lexical, morphological and syntactic transfer rules that translate Spanish input into Quechua sentences. In any such setup, certain ambiguities are extremely difficult to handle. For instance, if a word has more than one translation in the dictionary, a device for lexical selection is necessary in order to output the correct translation in the given context. This selection presents a great challenge for a rule-based architecture, since it is not feasible to cover all possible contexts with rules. Lexical disambiguation is thus one of the cases where we use statistical models in order to get the best possible output. This combination of rule-based and statistical approaches is generally referred to as *hybrid machine translation*.

In this chapter, we will describe our translation process from Spanish to Quechua step by step, from the analysis of the Spanish input, to the lexical and syntactic transfer, up to the generation of the Quechua output, see Figure 5.1 with a simplified illustration of the pipeline.

5.2 Analysis of Spanish Input

The first step in a rule-based translation system is the analysis of the input sentence in the source language, in this case Spanish. The quality of the translation will depend crucially on this first step: tagging or parsing errors will propagate and can eventually lead to a completely incomprehensible translation. As there are many tools for the analysis of Spanish text, we did not develop these modules ourselves but instead relied on already existing resources for this task.

Figure 5.2 illustrates the analysis in detail: the translation begins with sentence splitting, tokenization, named entity recognition and classification, and morphological analysis with the FreeLing libraries [Padró and Stanilovsky 2012]. Basically, FreeLing suggests all possible PoS tags for a given word form, but the decision as to which one is correct is left to the next module, a conditional random fields model trained with Wapiti [Lavergne et al. 2010]. The tagset contains not only information about the PoS, but also morphological features of a given word form, such as tense, mood and person for verbs

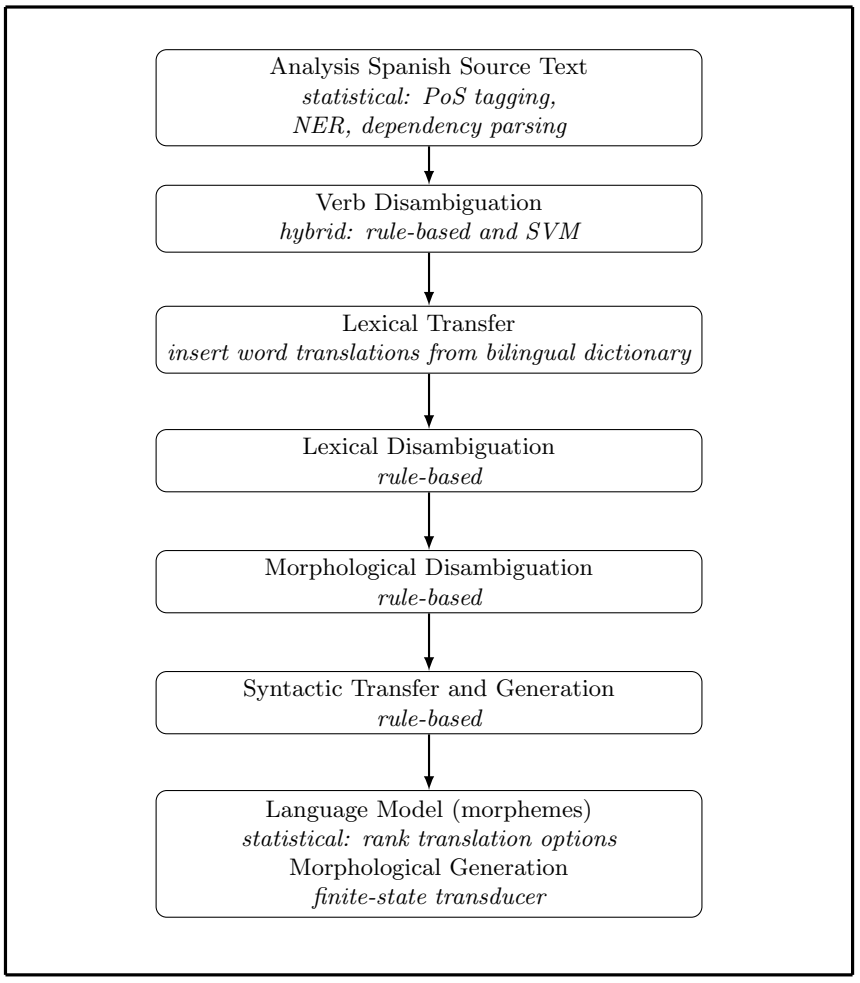


FIGURE 5.1: SQUOIA Translation Pipeline Spanish-Quechua

and gender, number and grade for nouns.[1] As loading the necessary FreeLing modules into memory is slow, we use a server-client architecture for this process. The FreeLing server receives the input text from the corresponding client and returns the analyzed text in the tabbed format needed for sequence labeling with Wapiti. The tagged output from Wapiti is then converted to CoNLL, the format used in the shared tasks of the Conference on Natural Language Learning[2], which serves as input to the dependency parser *DeSR* [Attardi 2006].[3] As loading the models for parsing into memory takes too

[1] A description of the Spanish EAGLES tagset is available at: http://nlp.lsi.upc.edu/freeling/doc/tagsets/tagset-es.html

[2] see http://ifarm.nl/signll/conll/

[3] In the current version, DeSR has been replaced with MaltParser, see http://www.maltparser.org/. However, since the evaluation presented in this chapter is based on the version with DeSR, we will focus on the original pipeline here.

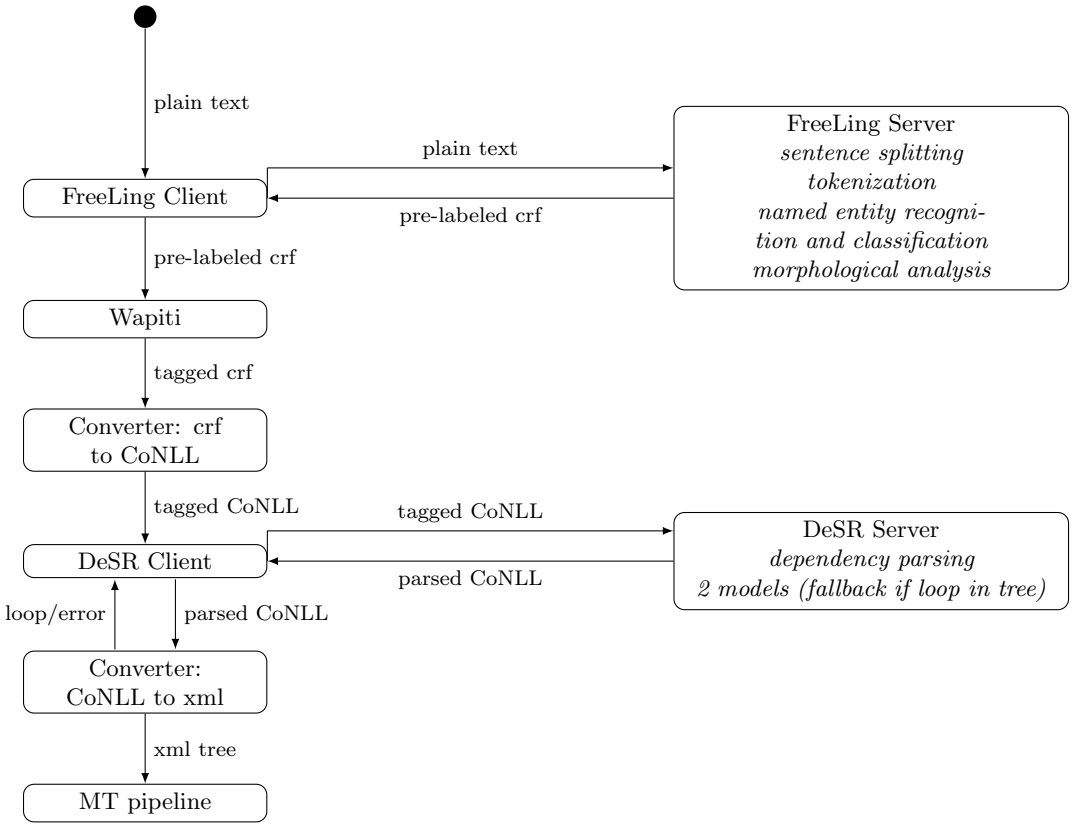


FIGURE 5.2: Analysis of the Spanish Input Sentences

much time to be efficient in the translation process, we also use a server-client implementation for parsing. The parser creates a dependency tree in CoNLL, but as this is not a suitable format for further processing, our next module will convert the parsed CoNLL into XML trees. During this conversion we insert chunks and also make some rule-based corrections of typical errors. For instance, if the parser labeled a noun as subject and its number does not match the number of the verb, we know for certain that this cannot be the subject of the clause, but instead is probably the direct object.[4] Furthermore, if the parser produced a loop or an unlikely tree (e.g. with a punctuation mark as head), the conversion module sends this sentence back to the parser, which in turn will re-parse the sentence using a different model. In the unlikely case that both statistical models produce a loop, the translation will stop at this point, as the system cannot create the XML tree.

The model for tagging with Wapiti was trained on a revised version of the Spanish

[4]Indirect and oblique objects are marked by a preposition in Spanish and are never mistaken for subjects by the parser.

treebank AnCora [Taulé et al. 2008]. The revision was necessary, as there were some inconsistencies between FreeLing, which suggests all possible tags for a word form, and AnCora. Furthermore, AnCora does not contain the whole tagset, as some rare word forms never occur in the corpus. Wapiti allows for pre-labeled input, which is useful if a word form has only one possible tag. However, if this pre-set label is unknown in the model, Wapiti treats it as a 'dummy' and overwrites it with one of the known labels. For this reason, we included a number of additional sentences that contained all the missing tags. Furthermore, we included parts of the biography of Gregorio Condori Mamani [Valderrama Fernandez and Escalante Gutierrez 1977] in order to have some typical Andean Spanish words covered as well: for instance, in AnCora, the word *papa* only occurs as a male noun ('the pope'), but in Andean texts, *papa* occurs most frequently as a feminine noun with the meaning 'potato' (peninsular Spanish: *patata*).

FreeLing offers two modules for tagging: one is a classical HMM (Hidden Markov Model) tagger, while the other one, called 'relax', relies on a combination of a statistical model with rules of a constraint grammar. Even though this hybrid approach achieves a better accuracy than the purely statistical tagger, our combination of FreeLing and Wapiti outperfoms both of the original FreeLing taggers by far as it reduces the error rate by half on most texts, see Table 5.1 for an evaluation on these four texts:

INFO *La papa y el cambio climático* - 'potatoes and climate change', inforesources 2008 (development aid, 456 sentences)[5]

COMERCIO *Lima la Brava* - 'Lima the wild one', newspaper column from *El comercio*, retrieved September 29, 2014[6]

WIKI first 4 paragraphs of the Spanish Wikipedia article about *Moneda* - 'Coin', retrieved September 29, 2014[7]

ROSABLANCA Andean story from the book *El hijo del oso* [Itier 2007]

Consider the Spanish sentence in example (50) with the corresponding output from FreeLing and Wapiti in Table 5.2.[1] Note that this is a made up sentence that includes as many ambiguities as possible that are addressed by specific modules during translation.

[5] Text from the SQUOIA treebank, see Chapter 3.

[6] http://elcomercio.pe/opinion/columnistas/lima-brava-gonzalo-torres-noticia-1760099

[7] http://es.wikipedia.org/wiki/Moneda

	Info	Comercio	Wiki	RosaBlanca
number of tokens	5625	574	1197	2646
FreeLing (HMM tagging)	94.44	92.74	93.92	95.08
FreeLing (relax tagging)	95.79	94.68	94.15	97.31
FreeLing (morph), Wapiti (tagging)	**98.76**	**97.58**	**98.48**	**97.74**

Table 5.1: Tagging Accuracy FreeLing and Wapiti

	tags suggested by FreeLing	unambiguous: tag assigned by FreeLing	ambiguous: tag assigned by Wapiti
Si	CS, NCMS000	-	CS
no	RN, NCMS000	-	RN
me	PP1CS000	PP1CS000	-
das	VMIP2S0	VMIP2S0	-
el	DA0MS0	DA0MS0	-
libro	NCMS000, VMIP1S0	-	NCMS000
que	PR0CN000, CS	-	PR0CN000
compré	VMIS1S0	VMIS1S0	-
en	SPS00	SPS00	-
Lima	NP00G00	NP00G00[8]	-
,	Fc	Fc	-
ya	RG, CS	-	RG
no	RN, NCMS000	-	RN
hablaré	VMIF1S0	VMIF1S0	-
contigo	PP2CSO00	PP2CSO00	-
.	Fp	Fp	-

Table 5.2: Morphological Analysis and Tagging with FreeLing and Wapiti

(50) *Si no me das el libro que compré en Lima, ya no hablaré contigo.*
if not me give2.SG.PRES the book that buy1.SG.PERF in Lima, anymore not speak1.SG.FUT with.you
'If you don't give me the book I bought in Lima, I won't talk to you anymore.'

The next step for the analysis of the Spanish input is dependency parsing with DeSR [Attardi 2006]. As mentioned above, we use two different models for parsing: the first model is the default, while the second works as a fallback, in cases where model 1 produces either a loop or a highly unlikely analysis (e.g. the root of the tree is a

[8] *Lima* is actually ambiguous, it could as well be a common noun ('file') or the 3rd person singular of the verb *limar* - 'to file'. However, as FreeLing has a module for named entity recognition and classification, we let FreeLing assign its tag (NP00G00: geographical proper name) instead of passing it to Wapiti.

id	word	lemma	s.tag	tag	morph.	head	label
1	Si	si	c	cs	_	4	conj
2	no	no	r	rn	_	4	mod
3	me	me	p	pp	gen=c|num=s|per=1	4	ci
4	das	dar	v	vm	gen=c|num=s|per=2|mod=i|ten=p	14	ao
5	el	el	d	da	gen=m|num=s	6	spec
6	libro	libro	n	nc	gen=m|num=s	4	cd
7	que	que	p	pr	gen=0|num=c	8	cd
8	compré	comprar	v	vm	gen=c|num=s|per=1|mod=i|ten=s	6	S
9	en	en	s	sp	gen=c—num=c—for=s	8	cc
10	Lima	lima	n	nc	gen=c|num=c|np=g0	9	sn
11	,	,	F	Fc	_	4	f
12	ya	ya	r	rg	_	14	cc
13	no	no	r	rn	_	14	mod
14	hablaré	hablar	v	vm	gen=c|num=s|per=1|mod=i|ten=f	0	sentence
15	contigo	contigo	p	pp	gen=c|num=s|per=2|cas=o	14	creg
16	.	.	F	Fp	_	14	f

TABLE 5.3: Dependency Parsing with DeSR (CoNLL)

punctuation mark). See Table 5.3 with the parsed sentence from example (50) in CoNLL format.[9] For better understandability, Figure 5.3 contains the graph representation of the CoNLL tree in Table 5.3.

The final step in the analysis of the Spanish source text is the conversion from CoNLL to XML. During this transformation, we apply several rule-based checks and corrections, additionally, chunks are created in order to have higher level units as possible targets for the translation rules, see Fig. A.1 in appendix A with the XML for example (50).

5.3 Verb Form Disambiguation

One of the most challenging parts for the translation from Spanish to Quechua is to determine which verb form to generate in subordinated clauses in the Quechua output.[10] There are two kinds of subordinated clauses that we need to disambiguate: clauses with a verbal head (complement clauses, final clauses, etc.) and clauses with a nominal

[9] s.tag = short tag, morph. = morphological features: these features contain the same information as the EAGLES tags from the previous step, see 5.2.

[10] Parts of this chapter are based on Rios and Göhring [2013] and Rios and Göhring [2016] (forthcoming).

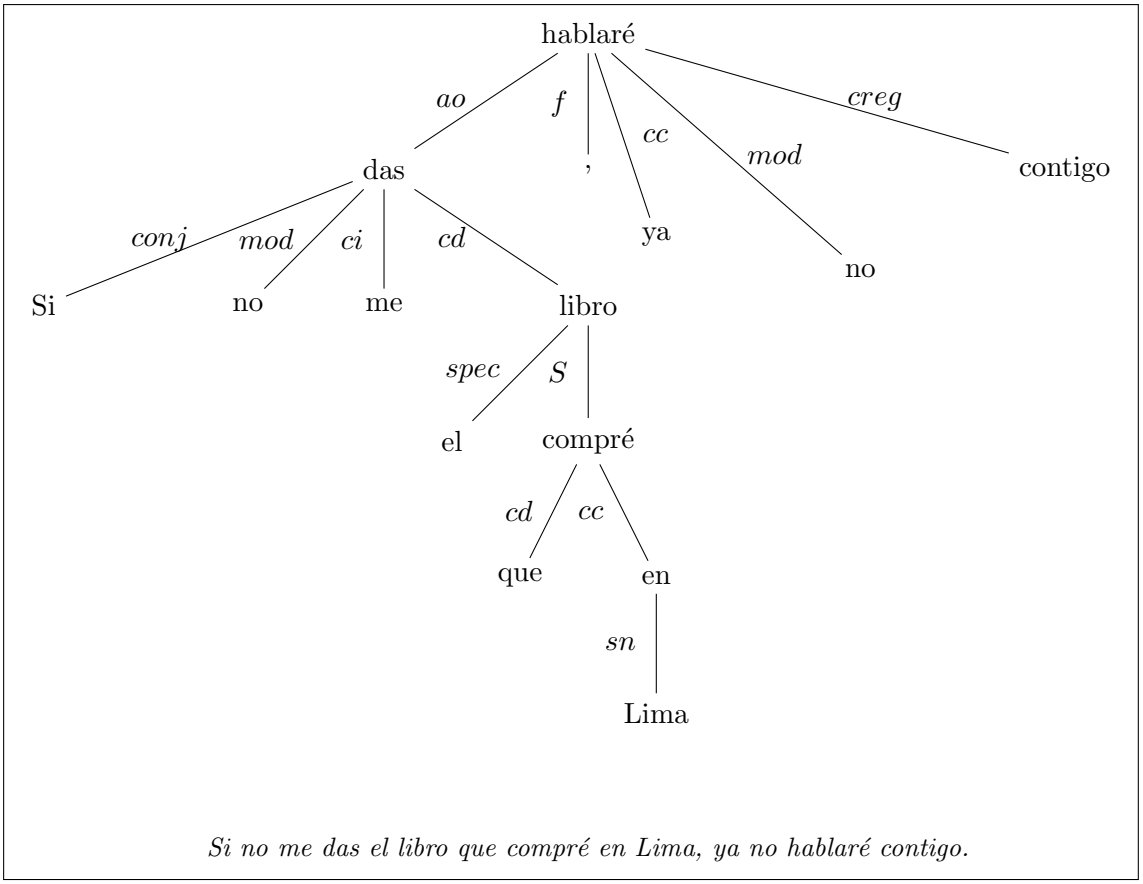


FIGURE 5.3: Spanish Dependency Tree according to CoNLL in Table 5.3

head (relative clauses). The former includes chaining by switch reference, in which case coreference resolution on subjects is necessary to translate the verb form correctly.[11].

5.3.1 Relative Clauses

Relative clauses in Quechua are nominal forms that are either agentive or non-agentive. For non-agentive relative clauses, there are two nominalizing suffixes available: *-sqa* ('perfect') is used for actions that have been completed, whereas *-na* ('obligative') occurs in contexts where the action has not been completed or indicates an intention, obligation or purpose. Consider the following examples:

[11]Switch-reference is a special type of clause linkage, where the subordinated verb has a marker that indicates if its subject is the same in the main clause. Switch-reference in Quechua will be explained in more detail in section 5.3.2 of this chapter.

(51) agentive: *Tunas mikuq pisqu takichkan.*

Tunas miku ***-q*** *pisqu taki -chka -n.*
tunas eat **-AG** bird sing -PROG -3.SG

'The bird that eats tunas is singing.'

(lit. 'the tunas-eating bird is singing') [Soto Ruiz 1976:153]

(52) non-agentive:

a. *Allin mikusqa kuchiqa wiram [kan].*

Allin miku ***-sqa*** *kuchi -qa wira -m.*
well eat **-PERF** pig -TOP fat -DIRE

'[A] well eaten pig [is] fat.' [Soto Ruiz 1976:153]

b. *Qamtaq, Gregorio, monta****na****y caballoyta hap'iy!*

Qam -taq, Gregorio, monta ***-na*** *-y caballo -y -ta*
you -CON Gregorio ride **-OBL** -1.SG.POSS horse -1.SG.POSS -ACC
hap'i -y!
grab -2.SG.IMP

'And you, Gregorio, grab my riding horse!'

(lit. 'grab the horse that I will ride/intend to ride')

[Valderrama Fernandez and Escalante Gutierrez 1977]

In order to generate the correct verb form for a Quechua relative clause, the first step is to automatically distinguish between relativization on subjects and relativization on obliques.[12] The latter are always translated with the non-agentive forms, but relative clauses where the head noun is the subject need to be further disambiguated: if the subject is a semantic agent, the verb in the relative clause has to be generated in the agentive form (*-q*), if the subject is not agentive, either *-sqa* or *-na* is the correct form, depending on tense, aspect and mood of the Spanish verb.

However, relative clauses in Spanish can be ambiguous, consider the following examples:

(53) agentive:

la mujer que comió el pan
the woman REL ate the bread

'the woman who ate the bread'

[12] Relativization on elements other than subject and object, e.g. with English *whose*.

(54) non-agentive:

el pan que comió la mujer
the bread REL ate the woman

'the bread that the woman ate'

The only difference between sentence (53) and (54) is the semantic class of the head noun: the verb *comer* - 'to eat' requires an animate, agentive subject like *mujer*. An inanimate noun like *bread* can therefore not be the subject of *comer*. The correct translation of example (53) uses the verb form with *-q*, whereas the verb in (54) should be translated with *-sqa*:

(55) agentive:

t'anta mikhu ***-q*** *warmi*
bread eat **-AG** woman

'the woman who eats/ate the bread'

(56) non-agentive:

warmi -p mikhu ***-sqa*** *-n t'anta*
woman -GEN eat **-PERF** -3.SG.POSS bread

'the bread that the woman eats/ate'

Not every Spanish relative clause is as ambiguous as the examples in (53) and (54). In the following cases, the head noun cannot be the subject of the relative clause, and therefore the agentive form can be discarded for the translation:

1. if the relative pronoun is preceded by a preposition (*el hombre* ***a quien*** *vió*),
2. if the relative pronoun is something other than *que, quien* or *cual*
3. if the verb in the relative clause is not congruent with the head noun
4. if the relative clause contains a subject noun or pronoun

However, case (4) is not a reliable feature in the translation process, as our parser frequently labels subjects as objects and vice versa, therefore, even if the parser detected a subject in the relative clause, the following disambiguation steps will still be applied.

Our rule-based module relies on a lexicon of Spanish verb frames [Taulé et al. 2008]: if the verb has only one frame, and the frame is intransitive, the head noun must be the subject. The semantic role indicated in the lexicon (agent, patient, impersonal, causer etc.) is the key to the correct translation: the Quechua verb should be rendered with the *-q* form, if the semantic role is agentive. In all other cases, the verb form in Quechua should be generated with either *-sqa* or *-na*. The decision as to whether the obligative or perfect form is correct depends on tense, aspect and mood of the Spanish verb.

If the frame retrieved from the semantic lexicon is transitive or ditransitive, the head noun is either the subject or object, but never the indirect object, as in this case the relative pronoun is preceded by the preposition *a*:

(57) indirect object as head of a relative clause:

el vecino ***a*** *quien la mujer muestra el libro*
the neighbor to REL the woman shows the book

'the neighbor, to whom the woman shows the book'

If the verb frame is transitive or ditransitive with an agentive subject, we cannot know whether the head noun is the subject or the object (see examples (53) and (54)). In case the verb lexicon contains more than one possible frame for a given verb, our module deletes all inapplicable frames with some additional context checks. If the frames cannot be reduced to one semantic role for the subject, the module takes a guess based on the semantics of the head noun. In this case, the disambiguation module retrieves the relevant information about the head noun from a semantic noun lexicon [Marimon et al. 2007]: if the head noun belongs to a semantic class that is a likely agent (e.g. animate, human, a social group), the module assigns the agentive form, but if the head noun is an unlikely agent (e.g. an inanimate or an abstract noun, a plant) it assumes that one of the non-agentive verb forms is correct.

The basic assumption is that only nouns of certain semantic groups are likely agents, while others are not (e.g. plants, abstract nouns, inanimates). This premise is of course

not always correct, therefore we tested a machine learning approach to disambiguate relative clauses.

5.3.1.1 Relative Clause Disambiguation with Machine Learning

The disambiguation of relative clauses with machine learning differs substantially from the disambiguation of other subordinated verb forms. Section 5.3.3.1 illustrates how our MT system relies on a classifier to determine the Quechua verb form in cases where the analysis of the Spanish source sentence went wrong. In the experiments with relative clauses, on the other hand, we use a classifier to assign the correct form instead of guessing the form based on semantic information in highly ambiguous cases.

5.3.1.2 Training Data

The training material for machine learning consists of relative clauses from the AnCora[13] and IULA[14] treebanks [Marimon et al. 2012, Taulé et al. 2008]. We let the rule-based module described in the previous section assign a form to all the relative clauses in the those treebanks and then extract those as training instances for the classifier. Most relative clauses in the treebanks are not ambiguous: as AnCora and IULA are manually annotated,[15] the annotation of subjects in relative clauses is reliable, as opposed to automatically parsed texts. In this case, we know that the relative clause is non-agentive: since the subject role is already occupied by an element of the relative clause itself, the head noun cannot be the subject. Furthermore, if the verb has only intransitive frames with either agentive or non-agentive subjects, we need no further disambiguation, as we can rely on the semantic role of the subject given in the verb frame lexicon. We manually checked and corrected all the ambiguous cases in AnCora and IULA that the rule-based module had to guess. Note that not all relative clauses are interesting for training, as we want to use the classifier only on ambiguous forms that cannot be determined by considering only the syntactic context. With this approach, we extracted 5,018 instances from AnCora and 3,201 instances from IULA to train the classifier.

[13] see http://clic.ub.edu/corpus/en/ancora

[14] see http://www.iula.upf.edu/recurs01_tbk_uk.htm

[15] IULA is parsed automatically, however the selection of the correct parse is done manually (see technical manual at http://iula05v.upf.edu/TreebankBrowser/docs/IULATreebank_TechnicalManual.pdf).

5.3.1.3 Features

We decided to use libsvm [Chang and Lin 2011] for the classification, as it provides a simple way of optimizing the parameters c (cost) and g (gamma) via grid search. In addition to the verb frames [Taulé et al. 2008] and the semantic noun classes of the Spanish Resource Grammar (SRG) [Marimon et al. 2007] used by the rule-based module, we integrated semantic information about the verb and the head noun from the Spanish wordnet [Gonzalez-Agirre et al. 2012] to the classification with libsvm. The semantic noun classes of the Spanish Resource Grammar are broad categories, such as *human*, *body part*, *plant* or *abstract noun*. The classes from the Spanish wordnet overlap with these in part, but are more fine-grained for abstract nouns, they include e.g. *feeling*, *event*, *phenomenon*, *motive*, *process* and some more.

Furthermore, we included some syntactic information, or more specifically whether the relative clause contains:

- the reflexive *se*[16]
- an indirect object
- a prepositional object
- an adjunct
- the demoted subject of a passive clause
- a predicative element (in equational clauses)

Note that we did not include the presence of a subject or direct object in the relative clause as features, as we cannot safely rely on the parser for this distinction.

Furthermore, we included an additional binary feature that indicates whether the lemma of the verb in Quechua is the copula *ka-*. The reason behind this feature is that relative clauses with *ka-* use the agentive form, even though on a semantic level, the head noun is not an agent. Relative clauses with the copula thus do not follow the general rule, see example (58).

[16]The Spanish reflexive *se* is a device of intransitivization and thus relevant for filtering out verb frames that do not match the context of a given relative clause.

(58) *urqukunapi kaq ayllukuna*

urqu -kuna -pi ka -q ayllu -kuna
mountain -PL -LOC be -AG village -PL

'mountain villages'

(lit. the villages that are in the mountains)

5.3.1.4 Evaluation

Our test set consists of 106 ambiguous relative clauses extracted from Spanish Wikipedia articles about three authors: Gabriel García Márquez, Mario Vargas Llosa and Pablo Neruda.[17]

The baseline in Table 5.4 is the performance of the rule-based module that guesses the form based on semantic information about the head. This simple guess was correct in 88 out of 106 cases, which results in 83.02% accuracy. As Table 5.4 shows, the SVM classifier does not achieve the accuracy of the rule-based method: even in the best setting, with all features, the classifier assigns the correct form only in 83 out of 106 cases. This results in an accuracy of 78.3%, which is worse than the performance of the rule-based module.

A possible explanation is the relatively small number of training instances: although we exploited two treebanks, the training set consists of only 8,219 instances. Furthermore, the training material is probably not as clean as the instances used for the disambiguation of the subordinated verbs in section 5.3.3.1: only the highly ambiguous (guessed) cases were manually checked, but there may also be a number of errors in the remaining relative clauses.

5.3.1.5 Relative Clauses with no Direct Correspondence

Not every Spanish relative clause can be directly translated into Quechua: for instance, relativizations with *cuya* - 'whose' and *donde* - 'where' do not have clear corresponding structures in Quechua:

[17] http://es.wikipedia.org/wiki/ retrieved 11.01.2014
[18] Settings: svm type=C-SVC, kernel type=RBF, cost=8, gamma=0.03125

	10x cross-validation	test set
features:[18]		
all features	**77.81**	**78.30**
no wordnet	75.46	75.47
no verb frames	72.89	64.15
no Resource Grammar noun classes	77.17	77.36
no syntactic features	76.10	75.47
baseline (rule-based, guessing)	–	**83.02**

TABLE 5.4: Evaluation of the SVM Classifier on Relative Clauses

(59) donde:

El pueblo donde yo vivo es grande.
the village where I live is big

'The village where I live is big.'

(60) cuyo/a:

La mujer cuyo hijo trabaja en la ciudad vive aquí.
the woman whose son works in the city lives here

'The woman whose son works in the city lives here.'

Example (59) can be expressed in Quechua with an internally headed relative clause: the system thus restructures the sentence by moving the head of the relative clause in front of the verb, while moving all chunk level morphology to the nominalized verb (topic marker *-qa* in this example). See Figure 5.4 with the resulting translation options for sentence (59) (*kawsa-* and *tiya-* are synonyms), example (61) provides the glosses to the system output in Fig. 5.4:

(61) El pueblo donde yo vivo es grande.

Ñuqa -p llaqta kawsa/tiya -sqa -y -qa hatun -mi [kan].
I -GEN village live -PERF -1.SG.POSS -TOP big -DIRE is

'The village I live in [is] big.'

```
ñuqap llaqta kawsasqayqa hatunmi. p:-22.2453
ñuqap llaqta tiyasqayqa hatunmi. p:-23.9417
```

Translation of: El pueblo donde yo vivo es grande.

FIGURE 5.4: Ranked Translation Options for Example (59)

Spanish relative clauses such as example (60) with *cuyo* have no corresponding form in Quechua: we would have to form 2 clauses to express this content ('The woman lives here, and her son works in the city'). However, this includes creating an entirely new syntax tree. In the current version, the SQUOIA system cannot handle this type of relative clause.

5.3.2 Coreference Resolution

A common type of subordination in Quechua is the so-called switch-reference: the subordinated, non-finite verb bears a suffix that indicates whether its subject is the same as in the main clause or not. If the subject in the subordinated clause is different, the non-finite verb additionally bears a possessive suffix that indicates the subject person.

(62) Same subject: *Mikhuspa hamuni.*

Mikhu ***-spa*** *hamu -ni.*
eat **-SS** come -1.SG

'When I finished eating, I come.'

(63) Different subject: *Mikhuchkaptiy pasakurqan.*

Mikhu -chka ***-pti*** *-y* *pasa -ku* *-rqa -n.*
eat -PROG **-DS** -1.SG.POSS leave -RFLX -PST -3.SG

'While I was eating, he left.'

[Dedenbach-Salazar Sáenz et al. 2002:168]

In the source language, Spanish, subordinated verbs are usually finite. An overt subject is not necessary, as personal pronouns are used only for emphasis ('pro-drop'). Consider the following example:

(64) *Cuando llegó a casa, María abrió la puerta.*
when arrive.3.SG.PST PREP house María open.3.SG.PST the door
'1. When she$_i$ came home, María$_i$ opened the door.'
'2. When she$_j$/he$_j$ came home, María$_i$ opened the door.'

In order to generate the correct verb form in Quechua, we need to know whether in Spanish the subject of the subordinated verb *llegó* is *María* or not. In the first case, we have to use the same subject form in Quechua, *-spa*, while in the second case, the correct form for the translation is *-pti*. As a preliminary solution to this problem, our system has a rule-based module that performs coreference resolution on subjects. So far, the procedure is based on the simple assumption that an elided subject is coreferential with the previous explicit subject, if this subject agrees in number and person with the current verb.

5.3.3 Disambiguation of Subordinated Clauses

Subordinated clauses in Quechua are often non-finite, nominal forms. There are several nominalizing suffixes that are used for different clause types that will be illustrated in more detail in this section.

Generally, the relation of the subordinated clause to the main clause is expressed through different conjunctions in Spanish. In Quechua, on the other hand, a specific verb form in combination with a case suffix indicates the type of subordination. For instance, Spanish *para que* - 'in order to' has to be translated with a nominal verb form with the suffix *-na* ('obligative') and the case suffix *-paq* (usually called benefactive, 'for'):

(65) *Ventanata kichay wayrap haykurimunanpaq.*

Ventana -ta kicha -y wayra -p hayku -ri -mu ***-na*** *-n*
window -ACC open -2.SG.IMP wind -GEN enter -INCH -DIR **-OBL** -3.SG.POSS
-paq.
-BEN

'Open the window for the air to come in!'
(lit. 'Open the window for his entering of the wind')

[Cusihuamán 1976:210]

Finite verb forms are also possible in subordinated clauses; in this case, the relation of the subordinated and the main clause is indicated through a 'linker'. A linker often consists of a demonstrative pronoun combined with case suffixes or so-called independent suffixes; these are special suffixes that can be attached to any word class and their position is usually at the end of the suffix sequence. The functions of the independent suffixes include data source, polar question marking and topic or contrast [Adelaar and Muysken 2004:209]. In combination with demonstrative pronouns, the independent suffixes are used for linking clauses similarly to Spanish or English conjunctions, see also the examples for the annotation of the treebank given in section 3.3.7.

For instance, the combination of demonstrative *chay* - 'this' with the topic marker *-qa*, *chayqa*, is used in the sense of 'if, in the case that':

(66) *Munanki **chayqa**, Arekipatapas rinki makinapi.*

*Muna -nki **chay -qa**, Arekipa -ta -pas ri -nki makina -pi.*
want -2.SG **this** **-TOP** Arequipa -ACC -ADD go -2.SG machine -LOC

'If you like, you can also go to Arequipa by train (machine).'

[Cusihuamán 1976:264]

Indirect speech in the Spanish source text is a special case, as the Quechua equivalence of indirect speech is direct speech. The conversion from indirect to direct speech is not trivial, because coreference resolution for the subject is required: if the subject of the main verb is the same as the subject of the indirect speech clause, the verb has to be generated as first person form in direct speech. Consider this English example:

(67) 'John said he wanted to go fishing.'

a. *if John = he:* 'I want to go fishing', John said.

b. *if John ≠ he:* 'He wants to go fishing', John said.

In this case, when both verbs have a 3rd person subject that matches in number, we naively consider both subjects as being equal and mark the direct speech Quechua verb as a first person form, as the current rule-based approach is unable to distinguish these two cases.

Furthermore, the form of the subordinated verb may also depend on the semantics of the main verb, e.g. complement clauses of control verbs usually require *-na* ('obligative'),

whereas with other verbs, the nominalizer *-sqa* ('nominal perfect') is used. With some verbs both forms are possible, generally *-na* implies that the action of the complement clause is not complete or has not started yet (also: intent, purpose), while *-sqa* means that the action has been completed:

(68) complement clauses:

a. *Ri* ***-na*** *-yki* *-ta* *muna -ni.*
go **-OBL** -2.SG.POSS -ACC want -1.SG
'I want you to leave.'
(lit. 'I want your going.')

b. *Ama -m* *chay yacha* ***-sqa*** *-yki* *-ta* *qunqa -nki* *-chu.*
don't -DIRE this know **-PERF** -2.SG.POSS -ACC forget -2.SG -NEG
'Don't forget what you learned.'
(lit. 'Don't forget that your learned.')

[Cusihuamán 1976:125]

For all of these cases, our translation system has a set of rules to match the given context, so that the correct form can be assigned to each verb.

5.3.3.1 Disambiguation of Subordinated Clauses with Machine Learning

5.3.3.1.1 Training Data

In order to generate the correct Quechua verb form in a subordinated clause, we need to extract the following information from the Spanish source sentence:

- semantics of the main verb
- the conjunction
- tense and mood of the subordinated verb (in some cases needed to distinguish between 'obligative' *-na* and 'perfect' *-sqa*)

Based on these features, the rule-based verb disambiguation module of our translation system assigns the Quechua verb form. Given a correct dependency tree, this rule-based approach achieves a high precision, but it fails if the parse tree contains errors. In order to obtain instances of main and subordinated clauses to train a classifier, we pre-translated

two manually annotated dependency treebanks: the Spanish AnCora dependency treebank[19] [Taulé et al. 2008] and the IULA Spanish LSP Treebank[20] [Marimon et al. 2012]. As these are correctly annotated, the rule-based module can disambiguate the subordinated verbs reliably, and we can extract these clauses as instances for training. With this approach, we collected 8,579 instances from AnCora and 5,704 from IULA[21], which results in a total of 14,283 instances for training.

5.3.3.1.2 Features

Instead of lemmas we use the semantic categories from the Spanish wordnet [Gonzalez-Agirre et al. 2012] and the AnCora verb frames [Taulé et al. 2008] to describe the main verb. For the subordinated verb, only tense and mood are relevant (extracted from the PoS tag in the treebank). For the conjunctions, we use the lexical forms, as there is no good way to describe them semantically. All features are binarized for training.

In our previous pipeline [Rios and Göhring 2013] we relied on the lemmas of main and subordinated verb instead of semantic and syntactic features. In this setting, Naïve Bayes achieved the best results. However, with the new set of features, the independence assumption[22] might not always be given, therefore we use support vector machines (libsvm) instead of Naïve Bayes.

5.3.3.1.3 Classification

Table 5.5 shows the accuracy of libsvm in 10-fold cross-validation and on a manually annotated test set of 100 instances. This is the same test set that we used before with Naïve Bayes [Rios and Göhring 2013]. For comparison, Table 5.5 also contains the results obtained with Naïve Bayes, once trained on the exactly same data set as libsvm, and

[19] `http://clic.ub.edu/corpus/en/ancora`

[20] IULA is parsed automatically, but the correct parse tree is selected manually. For more information, see `http://www.iula.upf.edu/recurs01_tbk_uk.htm`.

[21] Note that, although IULA contains more than twice as many sentences as AnCora, the sentences in IULA are mostly short, simple sentences, without subordinated clauses. The difference lies in the genre: IULA contains mostly abstracts from different scientific fields with a small contrastive corpus of newspaper articles, while in AnCora, all texts are newspaper articles.

[22] Classification with Naïve Bayes is restricted to cases where the value of each feature is independent from the value of all other features. If this condition is not met, Naïve Bayes will calculate wrong probabilities. With the new set of semantic features, we cannot safely assume that each feature is independent of the others, and for this reason, we cannot use Naïve Bayes with the new data.

Features	libsvm C-SVC, RBF, c=32, g=0.0078125		Naïve Bayes with semantic/ syntactic feat.		Naïve Bayes with lemmas	
	10x cv	test set	10x cv	test set	10x cv	test set
main verb, sub. verb, conjunction	**92.08**	**86**	81.47	75	84.28	78
sub. verb, conjunction	**87.97**	**81**	85.07	75	74.02	72

TABLE 5.5: Evaluation of the SVM Classifier on Subordinated Clauses[24]

once trained on the same data, but with verb lemmas instead of semantic and syntactic features. The results in Table 5.5 indicate that libsvm achieves the best accuracy, with 92.07% in cross-validation and 86% on the test set.[23]

The classification is slightly worse if only the conjunction and the subordinated verb are set, but the main verb is unknown (second line in Table 5.5). The third option, where the classifier has only information about the main and the subordinated verb while the conjunction is unknown, is not relevant: if no conjunction has been found, the module assumes that the verb form in question must be either a main verb, a relative clause or a coordination. All of these options are set by rules, not by the SVM classifier.

5.3.3.2 Rule-based Translation System with Machine Learning Verb Disambiguation

Figure 5.5 illustrates how the support vector machine (SVM) module is integrated into our translation pipeline: the rule-based verb disambiguation module tries to assign a Quechua form to all verbs in the Spanish tree. If the main verb or the conjunction is not found during this rule-based disambiguation, the verb form is marked as ambiguous and passed to the additional module for further disambiguation. This additional module checks in a first step if a given ambiguous verb form could be the actual main verb of the sentence or a relative clause that the parser attached to a non-nominal head. If this is the case, it assigns the verb form *finite* or *rel* for main or relative clauses respectively,

[23] In our previous setting with Naïve Bayes, we achieved only 81% accuracy, but we had a smaller training set of only ∼7,300 instances.

[24] C-support vector classification (C-SVC) with RBF kernel parameters c (cost) and g (gamma) obtained through search grid on 10-fold cross-validation (10x cv)

and the disambiguation is done. Otherwise, it checks if there is a conjunction: if there is one, the module looks for the main verb in the linear sequence of the tokens,[25] and then invokes the SVM model to assign a verb form. If there was no conjunction, the module assumes that this must be a coordination and assigns the same verb form as the preceding verb. If there is no preceding verb, this might be a tagging error (e.g. a noun that has been tagged as verb), in which case the module assigns the verb form *finite*, as this is the most common form.

5.3.3.3 Evaluation

5.3.3.3.1 Whole Verb Disambiguation Pipeline

We used the same four texts for the evaluation as in the previous setup [Rios and Göhring 2013]:

- *La catarata de la sirena* - 'the waterfall of the siren' (Andean story)
- the first two chapters of 'The Little Prince'
- an article from the Peruvian newspaper 'El Diario'
- the Spanish Wikipedia article about Peru

Since our previous publication [Rios and Göhring 2013], we have improved our tagger,[26] and therefore the number of recognized verbs is slightly higher than in the version from 2013. Our rule-based module disambiguates only 78.67% of all verb forms correctly, as it marks many verbs as ambiguous. In the next step, the additional disambiguation module with the SVM classifier assigns a verb form to all the ambiguous forms and thus increases the proportion of correct verb forms to 95.11%. The previous module, with Naïve Bayes, achieved only 89% accuracy on these texts, see Table 5.6.

5.3.3.3.2 Additional Verb Disambiguation Module

Furthermore, we used three longer texts to test the performance of the rule-based and the SVM part of the additional verb disambiguation module. As shown in Fig. 5.5, the

[25] The first verb to the left or right that is not an auxiliary and with no conjunction or relative pronoun between them.

[26] See section 5.2 about the analysis of the Spanish source text for the translation.

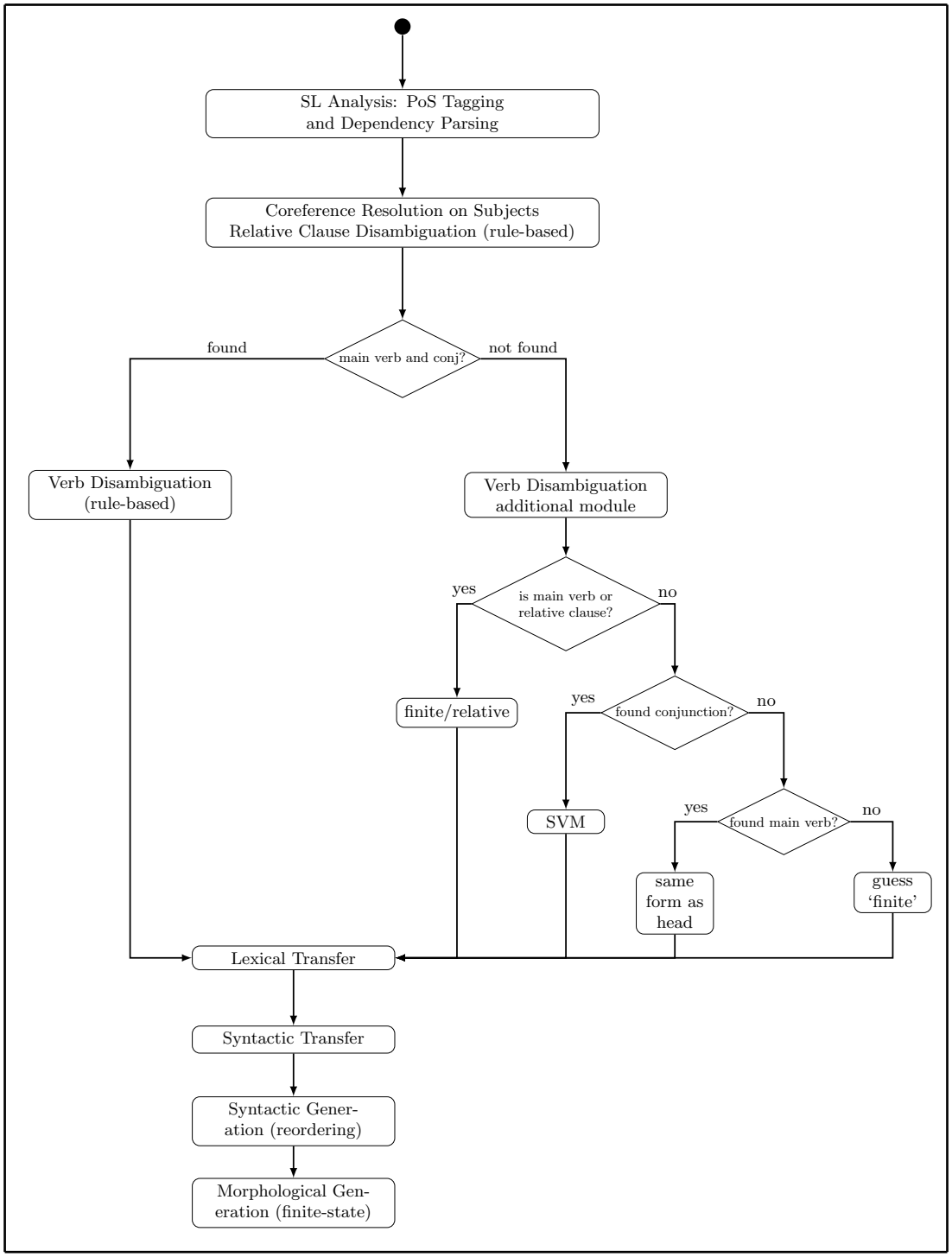


FIGURE 5.5: SVM Module in MT Pipeline

		correct	incorrect
rule based:	186	177	9
		78.67%	4.0%
with additional module (includes SVM) :	39	37	2
total 'verb' chunks:	225	214	11
		95.11%	**4.89%**
old version, with Naïve Bayes:		89.0%	11.0%

TABLE 5.6: Evaluation of Complete Disambiguation Pipeline

additional module relies on a set of rules to decide if the 'subordinated' verb in question is the actual main verb, a relative clause or a coordinated clause. If this is not the case, but the clause is clearly subordinated (indicated through a conjunction), the verb form is determined via SVM classification.

The texts that we used for this evaluation are part of the SQUOIA treebank (see Chapter 3):

- Festschrift 40th anniversary of the Peruvian-German chamber of commerce and industry (322 sentences)
- Memoria 2009, Peruvian-German chamber of commerce and industry (314 sentences)
- *La papa y el cambio climático* - 'potatoes and climate change', inforesources 2008 (development aid, 456 sentences)

Table 5.7 illustrates the performance of the additional verb disambiguation module. Most of the potentially ambiguous verbs (73 out of 92) are either main verbs, relative clauses or coordinations that the parser attached to the wrong head and were therefore not disambiguated by the rule-based module, but by the rule-based decision part of the additional verb disambiguation module. Not all ambiguous verb form candidates are actual verbs: the middle part of Table 5.7 shows 5 cases where nouns have been erroneously tagged as verbs. In total, the additional module assigned 79 out of 87 actual verb forms correctly, which results in 90.8% accuracy.

	rule-based decision (main verb, relative clause or coordination)	SVM	total
total ambiguous verb forms	73	19	92
total correct	64	15	79
			85.87%
total wrong	9	4	13
			14.13%
total tagging errors (no verbs)	4	1	5
total disambiguated (actual verbs)	69	18	87
correct	64	15	79
			90.8%
wrong	5	3	8
			9.2%

TABLE 5.7: Evaluation of the Additional Verb Disambiguation Module

Figure A.2 in appendix A illustrates the walkthrough example (50) from section 5.2 after the verb disambiguation: every `CHUNK` with `type="grup-verb"` or `type="coor-v"` has now an additional attribute `verbform`. The verb chunk with the main verb (*Hablaré*) is set to `verbform="main"`, which means the verb should be in finite form. Likewise, the verb in the subordinated clause, *das*, is marked as finite form, and additionally the conjunction *chayqa* is inserted: *chayqa* (`conjLast="chayqa"`) will be attached at the end of the conditional clause. Additionally, the verb in the relative clause, *compré*, was marked as non-agentive.

5.4 Lexical Transfer

The lexical transfer module reads the XML created by the conversion module described in section 5.2 and now annotated with verb forms, and inserts all possible translations for a given token from a bilingual dictionary, written in XML and compiled into a finite-state transducer for faster lookup.[27]

[27] We use the lttoolbox library from the Apertium platform to compile the dictionary, and a modified version of the lexical translation module from Matxin to insert the Quechua translations from the dictionary into the XML of the pipeline:
lttoolbox: `http://wiki.apertium.org/wiki/Lttoolbox`
Apertium: `http://wiki.apertium.org/wiki/Main_Page`
Matxin: `http://matxin.sourceforge.net/`

The core of our Spanish-Quechua lexicon consists of the Unified Southern Quechua dictionary by Cerrón-Palomino [1994], which contains ∼3500 nominal and ∼1700 verbal roots. As these numbers are too small to produce good translations, we enhanced the lexicon semi-automatically with entries from other sources, such as the dictionary of the AMLQ (*Academia Mayor de la Lengua Quechua*) in Cuzco. These additional resources are not written in the official Quechua orthography, therefore all new entries had to be manually checked and corrected. Furthermore, a native speaker from Cuzco translated 1000 verbs and nouns each. Our enlarged dictionary contains almost 85,600 Spanish lemmas, of which 18,738 have at least one translation to Quechua. Note that not all Spanish lemmas have a correspondence in Quechua, because for many non-native concepts, there is no native expression in Quechua. Nevertheless, it is beneficial to include the lemmas without translation into the dictionary, as morphological information will be inserted during the lexical transfer, depending on the paradigm a given word belongs to in the dictionary. This allows our system to produce a loan word in the Quechua output by attaching Quechua suffixes to the Spanish lemma in order to generate a complete word form.

Lexicalization patterns differ considerably between Spanish and Quechua, therefore the dictionary must contain specific information about the context for certain translations. Figure 5.6 illustrates the dictionary entry for the verb *transformar* - 'to transform': the Spanish verb is transitive ('to transform someone/something') in its unmarked form, but in Quechua, the unmarked form *tuku-* is intransitive. In Spanish, intransitivity is marked by the reflexive pronoun *se* (*transformarse*), and in Quechua, on the other hand, the transitive form is marked with the causative suffix *-chi* (*tukuchi-*). Both translations have an attribute `transitivity`, in this case with the values `trans` and `rflx` for the intransitive form with the reflexive pronoun. Both forms will be inserted during the lexical transfer, but depending on the context, i.e. the presence or absence of *se* in the Spanish clause, one of them will be deleted later. The second verb in Figure 5.6, *abrazar* - 'to hug', is transitive in Spanish as well, but when used with the reflexive pronoun *se* and a plural form, the meaning is reciprocal. In Quechua, reciprocity is expressed through the combination of *-na* (reciprocal) and *-ku* (reflexive), thus *marq'anaku-* - 'to hug each other'. Later on during the translation process, rules will determine if for a particular translation the unmarked or the reciprocal form suits best.

A special case in the dictionary is the Spanish verb *tener* - 'to have': there is no direct

```
<e>
 <p>
  <l>transformar</l>
  <r>tuku<s n="add_mi"/>+Caus<s n="transitivity"/>trans</r>
 </p><par n="Verb_main"/>
</e>
<e>
 <p>
  <l>transformar</l>
  <r>tuku<s n="transitivity"/>rflx</r>
 </p><par n="Verb_main"/>
</e>

<e>
 <p>
  <l>abrazar</l>
  <r>marq'a</r>
 </p><par n="Verb_main"/>
</e>
<e>
 <p>
  <l>abrazar</l>
  <r>marq'a<s n="add_mi"/>+Rzpr+Rflx<s n="transitivity"/>rzpr</r>
 </p><par n="Verb_main"/>
</e>
```

FIGURE 5.6: Lexicon Entry for *transformar* and *abrazar*

correspondence in Quechua, instead *tener* has different translations depending on the meaning in a given context, see Fig. 5.7 with some of the lexicon entries for *tener*. The default (and worst) translation is the last entry: 'unspecified' means that the Spanish root *tene-* will be used in the generation. However, the dictionary contains several translations for specific meanings (determined by the direct object), such as *tener hipo* → *hik'i-*, 'to have a hickup' or *tener vergüenza* → *p'inqa-*, 'to be ashamed'. Note that with some translations, such as *tener sueño* → *puñunaya-*, 'to be sleepy' (line 28 in Fig. 5.7), syntactic transformations are necessary, as the Spanish subject is realized as direct object in Quechua:

(69) 'to be sleepy':

a. *Tengo sueño.*
have.**1.SG** sleep
'I'm sleepy.'
(lit. 'I have sleep')

b. *Puñu -naya* ***-wa*** *-chka -n.*
sleep -DES **-1.OBJ** -PROG -3.SG
'I'm sleepy.'

[Soto Ruiz 2006:291]

```
<e>
 <p>
  <l>tener</l>
  <r>hik'i<s n="obj"/>hipo</r>
 </p><par n="Verb_main"/>
</e>
<e>
 <p>
  <l>tener</l>
  <r>ka<s n="obj"/>año,lugar</r>
 </p><par n="Verb_main"/>
</e>
<e>
 <p>
  <l>tener</l>
  <r>mancha<s n="add_mi"/>+Rflx<s n="obj"/>miedo</r>
 </p><par n="Verb_main"/>
</e>
<e>
 <p>
  <l>tener</l>
  <r>p'inqa<s n="add_mi"/>+Rflx<s n="obj"/>vergüenza</r>
 </p><par n="Verb_main"/>
</e>
<e>
 <p>
  <l>tener</l>
  <r>puñunaya<s n="subjToObj"/>1<s n="obj"/>sueño</r>
 </p><par n="Verb_main"/>
</e>
<e>
 <p>
  <l>tener</l>
  <r>unspecified</r>
 </p><par n="Verb_main"/>
</e>
```

FIGURE 5.7: Part of the Lexicon Entry for *tener*

Naturally, the correspondence between Spanish and Quechua lemmas is not always 1:1, but instead may be n:1, 1:n or even n:m. On the Spanish side, FreeLing relies on an list of multi-word expressions for tokenization and tagging. As FreeLing is a tool developed in Spain, the multi-word expression list included in the sources of the library contains predominantly peninsular locutions. We enhanced this list with some typical Peruvian locutions (e.g. *sacar la mugre* - 'to beat up', lit. 'to extract the filth/dirt'), but note that the list is far from complete. FreeLing will group these multi-word expressions into a single token during the analysis, which allows us to include a direct translation for these units in the bilingual dictionary.

In cases where a Spanish lemma corresponds to more than one Quechua word, the head in Quechua, usually the rightmost word, figures as translation, while the preceding words

are included as `preforms`: as suffixes will be attached exclusively to the head, we can already insert the `preforms` as complete word forms, see Fig. 5.8 with the following examples:

n:1

- *mayor de edad* - 'older': *kuraq*

1:n

- *forzar* - 'to force': *kallpawan kamachi-* (lit. 'to order/command with force')
- *memorizar* - 'to memorize': *umapi hap'i-* (lit. 'to grasp in the head')
- *concentrar* - 'to concentrate': *hunt'a sunqu yuya-* (lit. 'to think with full heart')
- *desempleo* - 'unemployment': *mana llamk'ayuq kay* (lit. 'the not-with-work-being')

n:m

- *estar en bola* - 'to be pregnant': *wiksayuq ka-* (lit. 'to be with belly')

Entries are organized in so-called `paradigms`, which are templates that group entries into classes and provide basic transfer information common to all translations within the paradigm. Figure 5.9 illustrates the template for main verbs with the tag `VMIP1S0` (V=verb, M=main, I=indicative, P=present, 1S=1.singular, 0=not a participle). Depending on the type of clause, the verb can have different forms: in a relative clause, the correct form is either `agentive`, `obligative` or `perfect`, whereas in a complement clause, `obligative` and `perfect` are possible. In a main clause or a finite subordinated clause, on the other hand, the correct form is `present`, while in a subordinated clause with certain conjunctions, such as *cuando* - 'when', the correct form is either `SS` (same subject) or `DS` (different subject). For a more detailed description of Quechua clause types and verb form disambiguation see section 5.3.

All morphological alternatives are inserted for every possible translation during the lexical transfer, see Fig. A.3 in appendix A with the XML of the first part of the example sentence (50) from section 5.2. Most of these lexical and morphological alternatives are wrong in the given context and need to be filtered out through morphological disambiguation.

```
n:1
<e>
 <p>
  <l>sacar_la_mugre</l>
  <r>maqa</r>
 </p><par n="Verb_main"/>
</e>
<e>
 <p>
  <l>mayor_de_edad</l>
  <r>kuraq</r>
 </p><par n="Noun"/>
</e>

1:n
<e>
 <p>
  <l>forzar</l>
  <r>kama<s n="add_mi"/>+Caus<s n="preform"/>kallpawan</r>
 </p><par n="Verb_main"/>
</e>
<e>
 <p>
  <l>memorizar</l>
  <r>hap'i<s n="preform"/>umapi</r>
 </p><par n="Verb_main"/>
</e>
<e>
 <p>
  <l>concentrar</l>
  <r>yuya<s n="preform"/>sunqu#hunt'a</r>
 </p><par n="Verb_main"/>
</e>
<e>
 <p>
  <l>desempleo</l>
  <r>kay<s n="preform"/>mana#llamk'ayniyuq</r>
 </p><par n="Noun"/>
</e>

n:m
<e>
 <p>
  <l>estar_en_bola</l>
  <r>ka<s n="preform"/>wiksayuq</r>
 </p><par n="Verb_main"/>
</e>
<e>
 <p>
  <l>diente_molar</l>
  <r>kiru<s n="preform"/>maran</r>
 </p><par n="Noun"/>
</e>
```

FIGURE 5.8: Lexicon Entries for 1:n, n:1 and n:m Translations

```
<pardef n="Verb_main">
<!-- 1.Sg.Present: e.g. ando, pongo, hago.. -->
 <e>
  <p>
   <l><s n="parol"/>VMIP1S0</l>
   <r><s n="mi"/>present<s n="verbmi"/>VRoot+1.Sg.Subj</r>
  </p>
 </e>
 <e>
  <p>
   <l><s n="parol"/>VMIP1S0</l>
   <r><s n="mi"/>SS<s n="verbmi"/>VRoot+SS</r>
  </p>
 </e>
 <e>
  <p>
   <l><s n="parol"/>VMIP1S0</l>
   <r><s n="mi"/>DS<s n="verbmi"/>VRoot+DS+1.Sg.Poss</r>
  </p>
 </e>
 <e>
  <p>
   <l><s n="parol"/>VMIP1S0</l>
   <r><s n="mi"/>agentive<s n="verbmi"/>VRoot+Ag</r>
  </p>
 </e>
 <e>
  <p>
   <l><s n="parol"/>VMIP1S0</l>
   <r><s n="mi"/>obligative<s n="verbmi"/>VRoot+Obl+1.Sg.Poss</r>
  </p>
 </e>
 <e>
  <p>
   <l><s n="parol"/>VMIP1S0</l>
   <r><s n="mi"/>perfect<s n="verbmi"/>VRoot+Perf+1.Sg.Poss</r>
  </p>
 </e>
...
</pardef>
```

FIGURE 5.9: Example Paradigm: Main Verbs

5.5 Morphological Disambiguation

Since many of the morphological translation options that were inserted during the lexical transfer are impossible in the given context, a special module filters the output according to hand-written rules. Every rule specifies a given context and whether to `keep` or `delete` the translation in question:

- `keep`: keep this translation, delete all others
- `delete`: delete this translation, keep others

```
1 source (Spanish) target keep/delete   context condition

3 my.smi=/^VMIS/  perfect     k  chunkparent.verbform=/rel:not\.agentive|perfect/
4 my.smi=/^VMIS/  obligative  k  chunkparent.verbform=obligative
5 my.smi=/^VMIS/  indirectpast,directpast  k  chunkparent.verbform=main

7 my.smi=/^VMIF/ future     k  chunkparent.verbform=main
8 my.smi=/^VMIF/ obligative k chunkparent.verbform=/obligative|rel:not\.agentive/
```

FIGURE 5.10: Rules for Morphological Disambiguation

The context for the rules can be established either with attribute-value pairs and simplified shortcuts for common XPath expressions, or through a full XPath expression for more elaborate contexts.

The context for the rules can be established either with attribute-value pairs and simplified shortcuts for common XPath expressions, or through a full XPath expression for more elaborate contexts. Figure 5.10 contains the rule to disambiguate the verb *compré* in the relative clause from the walkthrough example on page 85[28]. `chunkparent` is an abbreviation for the XPath expression `ancestor::CHUNK[1]`, the first ancestor that is a `CHUNK`. For *compré*, the rule in line 3 will be applied, as the verb chunk of the node *compré* was labeled as `rel:not.agentive`. Our system will keep the morphology to generate the perfect form of the verb and discard all others. If the assigned translation for a past form is finite, there is a further ambiguity: Quechua has two past forms, the narrative past with the suffix *-sqa*, which implies indirect evidentiality (speaker did not witness or experience the event), and a 'neutral past' with the suffix *-rqa*, that usually implies direct evidentiality, but see Faller [2004] for a detailed analysis of Quechua past forms and evidentiality. The system has no automatic method to decide if a given Spanish sentence should be translated with direct or indirect evidentiality as this would require information that is not present in the Spanish source text. For this reason, evidentiality is treated as a parameter that needs to be indicated to the system. If evidentiality was not set at start, the system will use direct evidentiality as default.

For the main verb *hablaré* with future tense in Spanish, the rule in line 7 of Fig. 5.10 will be applied: all translations except the one with the finite future form will be discarded. See Figure A.4 with the disambiguated XML: all verb chunks contain only one form at this point.

[28] *Si no me das el libro que* ***compré*** *en Lima, ya no hablaré contigo.*

Another important step for the morphological disambiguation is the translation of the Spanish prepositions. Quechua has no prepositions, but instead uses case suffixes or postpositions. A Spanish preposition can have more than one possible translation, depending on the context. For instance, the Spanish preposition *de* can have the following translations in Quechua:

- ablative suffix: origin or material (*Viene* ***de*** *Perú.* - *Peru suyu****manta*** *hamun.* / *Construyen la casa* ***de*** *ladrillos.* - *Ladrillu****manta*** *wasita ruwachkanku.*)
- genitive suffix: possession (*la casa* ***de*** *Juan* - *Juan****pa*** *wasin*)
- accusative suffix: with some verbs (*me acuerdo* ***de*** *ti* - *qam****ta*** *yuyarisayki*)
- locative suffix: temporal expressions (*el 28* ***de*** *octubre* - *28 punch'aw Kantaray killa****pi***)
- deleted: with materials (*casa de piedra* - *rumi wasi*)
- postposition *hina*: with some verbs (*ejercer* ***de****, hacer* ***de***)

Complex prepositions, i.e. prepositions with more than one word such as *debajo de* ('below') are treated as multi-word expressions by FreeLing and have their own lexicon entry and translation rules in our sytem. Especially with positional statements, the rules will also add a genitive case suffix to the nominal argument of the preposition. For instance, consider example (70), *delante de la casa* ('in front of the house'): *delante de* is translated as *ñawpan*, a postposition, and the nominal argument, *wasi* receives a genitive suffix *-p*. A literal translation of the Quechua phrase would be 'of the house in/at its front'.

(70) *wasi -p ñawpa -n -pi*
house -GEN front -3.SG.POSS -LOC
'delante de la casa' ('in front of the house')

(71) *mesa -p pacha -n -pi*
table -GEN floor -3.SG.POSS -LOC
'debajo de la mesa' ('below the table')

(72) wasi *-p hawa -n -pi*
house -GEN surface -3.SG.POSS -LOC
'encima de la casa' ('on top of the house')

[Cusihuamán 1976:140]

Figure A.4 in appendix A illustrates the XML for example (50) after the morphological disambiguation and insertion of the translations for the prepositions: the preposition *en* in ***en*** *Lima* has been translated as locative suffix (`case="+Loc"`). In all nodes with morphological ambiguities, the values of the correct SYN node have been copied to the NODE itself, while all the other SYN nodes have been discarded.

5.6 Syntactic Transfer and Generation

In order to obtain a grammatical Quechua sentence, certain pieces of information have to be moved within the XML tree, some additional information might be inserted and some nodes or even chunks will be deleted. Syntactic transfer in SQUOIA is handled by two separate modules: in a first step, one of the modules copies information from nodes within the chunk to the chunk itself or vice versa (intra-chunk transfer), whereas the second module moves information from chunk to chunk (inter-chunk transfer).

For instance, deontic *tener que* and *deber* ('must, have to') need a complete restructuring of the clause when translated to Quechua:

(73) *Tengo que lavar mi ropa.*
have.1.SG COMP wash my clothes
'I must wash my clothes.'

P'acha -y -mi t'aqsa -ku -na -y ka -chka -n.
clothes -1.SG.POSS -DIRE wash -RFLX -OBL -1.SG.POSS be -PROG -3.SG

'I must wash my clothes.'

[Cusihuamán 1976:210]

See Fig. 5.11 with the XML output for the sentence in example (73) after the syntactic transfer: first of all, the finite verb *tengo* is set to be omitted from the output, and since the attribute `deletefiniteMiInAux` is set, the system will not use the morphological information from the auxiliary (`VRoot+1.Sg.Subj`), but instead use the morphology provided in `addverbmi` (`+Obl+1.Sg.Poss`). For the main verb, *lavar*, additional morphology (`addverbmi`) is inserted: `+Obl` for the obligative suffix *-na*. Based on the tag of *tengo*, `VMIP1S0`, the person of the main verb is set to 1st singular, and as obligative

Tengo que lavar mi ropa.

```
<?xml version="1.0" encoding="UTF-8"?>
<corpus evidentiality="direct">
  <SENTENCE ref="1">
    <CHUNK ref="3" type="grup-verb" si="top" verbform="main" lem2="ka"
     addverbmi="+Obl+1.Sg.Poss" deletefiniteMiInAux="1"
     verbmi2="+3.Sg.Subj">
      <NODE ref="3" slem="lavar" smi="VMN0000" sform="lavar"
     UpCase="none" lem="muqch'i" mi="infinitive" verbmi="VRoot+Inf">
        <SYN lem="muqch'i" mi="infinitive" verbmi="VRoot+Inf"/>
        <SYN lem="t'aqsa" mi="infinitive" verbmi="VRoot+Inf"/>
        <SYN lem="maqchi" mi="infinitive" verbmi="VRoot+Inf"/>
        <SYN lem="maylla" mi="infinitive" verbmi="VRoot+Inf"/>
        <SYN lem="arma" mi="infinitive" verbmi="VRoot+Inf"/>
        <SYN lem="upha" mi="infinitive" verbmi="VRoot+Inf"/>
        <NODE ref="2" alloc="" slem="que" smi="CS" sform="que"
     UpCase="none" unknown="transfer">
          <NODE ref="1" slem="tener" smi="VMIP1S0" sform="Tengo"
     UpCase="none" lem="unspecified" mi="present"
     verbmi="VRoot+1.Sg.Subj"/>
        </NODE>
      </NODE>
      <CHUNK ref="5" type="sn" si="cd" poss="++1.Sg.Poss">
        <NODE ref="5" alloc="" slem="ropa" smi="NCFS000" sform="ropa"
     UpCase="none" lem="p'acha" mi="NRoot">
          <NODE ref="4" alloc="" slem="mi" smi="DP1CSS" sform="mi"
     UpCase="none" lem="" mi="++1.Sg.Poss"/>
        </NODE>
      </CHUNK>
      <CHUNK ref="6" type="F-term" si="term">
        <NODE ref="6" alloc="" slem="." smi="FP" sform="."
     UpCase="none" unknown="transfer"/>
      </CHUNK>
    </CHUNK>
  </SENTENCE>
</corpus>
```

FIGURE 5.11: Deontic *tener que* after Intra- and Interchunk Syntactic Transfer

is a nominal form, it must be a possessive suffix, not a verbal subject marker. Furthermore, the Quechua copula *ka-* is inserted as second lemma (`lem2`) with its own morphology (`verbmi2`). If the contents of `verbmi` in `NODE` (inserted during lexical transfer) and `addverbmi` in `CHUNK` (inserted during syntactic transfer) contradict each other, `addverbmi` will always prevail over `verbmi`.

Apart from restructuring the verb chunk, the syntactic transfer has moved the morphological information (`mi=++1.Sg.Poss`) from the possessive pronoun *mi* in the phrase *mi ropa* - 'my clothes' up into the nominal chunk. This is necessary, as the Spanish possessive pronoun corresponds to a possessive suffix in Quechua.

Figure A.5 in appendix A illustrates the XML after syntactic transfer for example (50):

the negation has been added to both clauses (`chunkmi="+Neg"`), furthermore, the object marker has been included into the verb chunk with the translation of *das* in the subordinated clause (`addverbmi="++1.Sg.Obj"`), while the pronoun *mi* itself is set to be deleted. Additionally, the tag for the suffix *-ña*, `+Disc`, has been included to the morphology of the negation particle *mana*. Finally, the tag for the case suffix *-ta*, `+Acc`, has been inserted into the nominal chunk with the translation of *book*, since this is the direct object of the verb *das*.

Once the necessary syntactic information is in place, the individual elements of the syntax tree need to be reordered according to the word order of the target language. The system has two modules for reordering: intra-chunk (reorders nodes in chunk) and inter-chunk (orders chunks). Both modules rely on a set of rules that define the correct output order.

Quechua is a head-final language, although word order in the main clause is relatively free [Sánchez 2010:12]: the subordinated clause precedes the main clause, and attributive nouns, adjectives and relative clauses precede the noun they modify. Furthermore, Spanish prepositions correspond to Quechua suffixes or postpositions, and clauses are generally verb-final. Every chunk receives an attribute `ord` that defines its position in the final output. If a chunk contains more than one node, the nodes are ordered relative to each other.

5.7 Ranking and Morphological Generation

As not all ambiguities can be resolved by rules, the SQUOIA translation system may end up with several different translation options for a given Spanish sentence. In this case, the system needs a device to find the best translation, or the n-best translations. A statistical language model provides exactly this functionality: language models are commonly used in statistical MT systems in order to assess the probability that a given translation is a fluent sentence in the target language [Koehn 2010:181ff.]. In SMT systems, this includes decisions on lexical choices as well as variations in word order, while for the SQUOIA translation system the latter is irrelevant, as word order follows a fixed template defined by rules, see section 5.6. The language model in the SQUOIA system thus mainly helps with the lexical choices.

```
NP suyu +Gen hatun rima +Rzpr +Rflx +Inf +3.Sg.Poss , acuerdo nacional ni +Perf
... kuska +Lim kawsa +Inf wiña +Inch +Caus +Obl +Ben NP suyu rima +Rzpr +Rflx
     +Inf +3.Sg.Poss +Top wiraqucha
NP NP NP , presidente
constitucional ni +Perf mink'a +Rflx +Perf +3.Sg.Poss +Instr +DirE
qillqa +Inch +Rflx +3.Sg.Subj.IPst 22 anta sitwa killa +Loc , 2002 wata +Loc .
```

Text version:
Perú Suyupa Hatun Rimanakuynin, Acuerdo Nacional nisqa
...Kuskalla kawsay wiñarichinapaq Perú Suyu Rimanakuyninqa wiraqucha
Alejandro Toledo Manrique, Presidente
Constitucional nisqa minkakusqanwanmi
Qillqarikusqa 22 anta sitwa killapi, 2002 watapi.

FIGURE 5.12: Stems and Morphemes in Training Corpus for Language Model

Language models are built on monolingual texts, ideally on large corpora. Written Quechua texts are still relatively rare: apart from books with Andean stories and translations of religious texts, material in Quechua is scarce. A further complication for statistical methods is the lack of a widely used and accepted standard orthography (see section 2.2).

The training corpus for the language model used in SQUOIA amounts to ∼58,000 sentences and consists of several printed texts that we scanned and manually corrected, but also several official Peruvian documents, such as the constitution. The largest part of the corpus, however, is a translation of the Bible (∼41,000 sentences). It is important to note that this translation was done by several people, each writing in their own dialect and orthography, and some parts are translated very poorly (ungrammatical word forms, typos, etc.), but due to the lack of other texts, we had to include the Bible.

All texts were automatically normalized through the pipeline described in section 2.3, and the remaining ambiguities were manually resolved. The language model that we trained on this corpus provided very low probabilities on our machine translated texts, as many words were unknown to the model, often due to small differences in the suffix sequence in otherwise known word forms. Therefore, we decided to build the language model on morphemes instead of words. We also extracted semi-automatically a list of all the proper names in our corpus and replaced them in the texts with `NP` (*nombre propio*), in order to avoid especially the numerous names in the Bible that will probably never occur in other texts. Figure 5.12 illustrates a short excerpt from the training corpus as opposed to the original text, a translation of the *Acuerdo Nacional 2002* [Chávez Gonzales et al. 2002].

```
mana:+DirE
Lima:+Loc
ranti:+Perf+1.Sg.Poss
libro:+Acc
qu:+1.Sg.Obj+2.Sg.Subj+Neg
chayqa:
,-PUNC-Fc
mana:+Disc+DirE
qam:+Instr
rima:+1.Sg.Subj.Fut+Neg
.-PUNC-Fp
#EOS


Manam Limapi rantisqay librota quwankichu chayqa , manañam qamwan
    rimasaqchu . p:-67.6535
```

Si no me das el libro que compré en Lima, ya no hablaré contigo.

FIGURE 5.13: Translated Output for Example (50)

The input for the application of the language model in our MT system consists of the stems[29] and tags that the translation system produces during the transfer. Once the n-best translations have been identified, the system calls a finite-state transducer to generate the word forms based on the stems and tags. As we already know which translation is the best (or n-best), we do not generate the options with lower probabilities. For the ranking we use kenlm [Heafield 2011], the morphological generation relies on a finite-state transducer implemented in foma [Hulden 2009a]. As both tools provide a C/C++ API, a single module ranks the options and then generates the n-best translations. A further advantage of this approach besides the sparse data reduction is that we only generate the n-best translations and discard the rest, as opposed to the ranking on word forms, where we have to fully produce all translation options in order to rate them. Ranking on morphemes thus saves computation time for sentences with lexical ambiguities.

Figure 5.13 contains the morphological output of the sentence in example (50) and the final translation generated with the foma finite-state transducer. As this sentence was fully disambiguated, there is only one translation option. For the sentence from example (73) on the other hand, *Tengo que lavar mi ropa*, several options are possible, as *lavar*

[29] Stem here refers to the combination of the root and an optional suffix from nominal or verbal slot 1, see Table 2.6 and 2.7 in Chapter 2. We do not split most of these suffixes: since they affect the lexical translation from Spanish, they are included in the lexicon entries of the dictionary and are never independently inserted during the translation process. The only exceptions are the reflexive *-ku* and the inchoative *-ri*.

```
p'acha:+1.Sg.Poss
muqch'i:+Obl+1.Sg.Poss
/t'aqsa:+Obl+1.Sg.Poss
/maqchi:+Obl+1.Sg.Poss
/maylla:+Obl+1.Sg.Poss
/arma:+Obl+1.Sg.Poss
/upha:+Obl+1.Sg.Poss
ka:+3.Sg.Subj
.-PUNC-FP


P'achay t'aqsanay kan. p:-16.7188
P'achay armanay kan. p:-16.8711
P'achay mayllanay kan. p:-18.3904
P'achay maqchinay kan. p:-18.9452
P'achay muqch'inay kan. p:-19.0055
P'achay uphanay kan. p:-19.0055
```

Tengo que lavar mi ropa.

FIGURE 5.14: Ranked Translation Options for Example (73)

-'to wash' has different translations in Quechua, depending on the object that undergoes the washing:

- *t'aqsa-* 'to wash textiles or hair'
- *arma-* 'to wash the body, bathe'
- *maylla-* 'to wash hands or feet'
- *maqchi-* 'to rinse'
- *muqch'i-* 'to wash/rinse mouth'
- *upha-* 'to wash the face'

Figure 5.14 illustrates the morphological output of the system and the ranking of all translation options. As expected, the translation *t'aqsa-* 'to wash textiles' is the best translation in the context of washing clothes.

5.8 Discourse: Modeling Topic and Focus

A special feature of Quechua, as opposed to the mainstream languages dealt with in machine translation, is the morphological marking of discourse-relevant information: the placement of evidential, negation and interrogative suffixes depends on the information structure of the sentence, since they are usually attached to the focalized element. In

contrast, Quechua has two suffixes, *-qa* and *-ri*, that function as topic markers. Furthermore, alternations in the information structure of a Quechua sentence can be expressed through a change of word order. Since information structure can be indicated by morphological markers on *in situ* elements [Sánchez 2010:30], a change in word order is however not mandatory. Section 5.8.1 describes the usage of the Quechua discourse morphology and some of the known restrictions as to the placement of the relevant suffixes, while section 5.8.2 presents different approaches to include topic and focus markers in the output of the translation system.

5.8.1 Discourse Morphology and Information Structure in Quechua

Southern Quechua has two topic markers, *-qa* and *-ri*, that occur in the same slot and are mutually exclusive. Following the definition in Sánchez [2010:43], *-qa* is the morphological marker that identifies information that is discourse-accessible and a matter of common concern to speaker and addressee. The suffix *-ri* occurs in questions and links the question to a prior event or statement, and as such motivates the addressee to continue the conversation [Cusihuamán 1976:227]. However, there seems to be some dialectal variation as to the usage of *-ri* and *-qa*: one of the texts in the SQUOIA treebank that was translated by a native speaker of Moquegua Quechua contains numerous sentences where *-ri* is used in declarative clauses instead of *-qa*.[30] A clause may contain more than one *-qa*, and it can only be attached to full clause constituents, however. Furthermore, *-qa* is restricted to finite clauses, since constituents inside a nominalized subordinated clause cannot be topicalized [Sánchez 2010].

Focus marking in Quechua depends on the clause type: in declarative, non-negated clauses, the evidential suffixes mark focus, see examples (74)-(76). In negation, the evidential is attached to the particle *mana* ('no'), while the negation suffix *-chu* takes over the function as focus marker, see examples (77)-(78). Similarly, in yes-no questions without negation, the interrogative suffix *-chu* marks the focalized part of the clause. In questions with interrogative pronouns (*wh*-words), the interrogative pronoun is marked

[30] Translation of the annual report 2009 of the Deutsche Welle Academy about *Development and the Media* see `http://www.dw.de/`. The Quechua version is available from the SQUOIA repository at `https://github.com/ariosquoia/squoia/blob/master/treebanks/texts/quz/DW_qu.txt`.

for either contrastive focus with the suffix *-taq*, or with an evidential, see examples (79)-(82).

declarative affirmative clauses:

(74) *Allqu -m kawallu -ta kani -n.*
dog -DIRE horse -ACC bite -3.SG
'It's the dog that bites the horse.'

(75) *Kawallu -ta -m allqu -qa kani -n.*
horse -ACC -DIRE dog -TOP bite -3.SG
'It's the horse that the dog bites.'

(76) *Kani -n -mi kawallu -ta allqu -qa.*
bite -3.SG -DIRE horse -ACC dog -TOP
'The dog BITES the horse.'

[Sánchez 2010:47]

declarative negated clause:

(77) *Mana -m allqu -chu kawallu -ta kani -n.*
not -DIRE dog -NEG horse -ACC bite -3.SG
'It's not the dog that bites the horse.'

(78) *Mana -m kawallu -ta -chu allqu -qa kani -n.*
not -DIRE horse -ACC -NEG dog -TOP bite -3.SG
'It's not the horse that the dog bites.'

yes-no questions:

(79) *Kay -pi -chu tiya -nki?*
this -LOC -INTR live -2.SG
'Do you live here?'

[Cusihuamán 1976:259]

(80) *Mana -chu miku -nki?*
Not -INTR eat -2.SG
'Do you not eat?'

[Sánchez 2010:79]

wh-questions:

(81) *Ima -ta -m muna -nki?*
what -ACC -DIRE want -2.SG
'What do you want?'

[Cusihuamán 1976:257]

(82) *May -manta -taq ka -nki -ri?*
where -ABL -INTR be -2.SG -TOP
'And where are you from?'

[Cusihuamán 1976:258]

There are several restrictions on the placement of the evidential suffixes that mark focus, according to Sánchez [2010:60]:

1. they are constituent-external
2. they are limited to one per clause
3. they can only appear in main clauses or in subordinated clauses with tensed verbs, not in nominalized subordinated clauses
4. they cannot occur in imperatives
5. they cannot occur in gapping expressions
6. they cannot be attached to modifiers that are unmarked for case

Table 5.8 contains the number of topic and focus markers per clause in the SQUOIA treebank. The numbers reveal that there is indeed a strong tendency for a clause to contain only one focus marker. There are however 36 sentences that violate restriction 2, consider the following example (taken from Gregorio Condori's autobiography):

(83) *Chay manka -kuna -ta* ***-m*** *San Pedro -pi* ***-m*** *ranti -rqa -ni,*
this pot -PL -ACC **-DIRE** San Pedro -LOC **-DIRE** buy -PST -1.SG
hina -spa -m apa -mu -rqa -ni qhiswa sara -wan chhala -na
be.like -SS -DIRE take -DIR -PST -1SG qhiswa rice -INSTR exchange -OBL
-y -paq.
-1.SG.POSS -BEN
'These pots, I bought [them] in San Pedro and I took them [there] to exchange [them] with qhiswa rice.'

[Valderrama Fernandez and Escalante Gutierrez 1977]

The main verb *rantirqani* has two constituents marked for direct evidentiality, the direct object *mankakunata* and the location *San Pedropi*. However, clauses such as example (83) are rare, most cases of double focus in the treebank are either coordinations where both elements are marked, or subordinated clauses with linkers such as *hinaspam/hinaspas* (see example (83) for glosses) and *chaymi/chaysi* (*chay* - 'this' and *-mi/-si* evidentials). This could indicate a certain degree of lexicalization as clause linkers for these forms, where the evidential suffixes no longer function as markers on their own, see example (84):

(84) *..papa -manta imaymana mikhu -na -kuna -qa hatarpari -n, chay* ***-mi***
potato -ABL all.kinds eat -OBL -PL -TOP rise -3.SG this **-DIRE**
papa -qa achkha -ta -puni ***-m*** *tarpu -ku -n.*
potato -TOP much -ACC -DEF **-DIRE** SOW -RFLX -3.SG
'..[consumption of] all kinds of food [made] from potato has increased, therefore much potato is sowed.'[31]

Overall, clauses with more than one focalized element are rare and can be ignored for the output of the machine translation: our system adheres to restrictions (2) and (3) from above, thus producing only sentences where focus is limited to one per finite clause and focus markers do not occur within nominalized clauses.

A special case for translation is that of equational clauses with the copula *ka-*, since there is evidence from our treebank that the distribution of topic and focus markers is relatively fixed: there is a strong tendency for the subject to be marked as topic and for

[31] From the translation of *La papa y el cambio climático* - 'potatoes and climate change', inforesources 2008.

clauses with 1 topic child:	1086
clauses with 2 topic children:	195
clauses with 3 topic children:	15
clauses with 4 topic children:	1
clauses with >4 topic children:	0
clauses with 1 focus child:	1322
clauses with 2 focus children:	36
clauses with >2 focus children:	0
total number of sentences:	1979

TABLE 5.8: Number of Topic and Focus Markers per Clause in the SQUOIA Treebank

the predicative element to be marked for focus, see Table 5.9. A typical example of this is (85), but also examples (86) and (87):

(85) *Qam -qa allin warmi -m ka -nki.*
you -TOP good woman -DIRE be -2.SG
'You are a good woman.'

[Cusihuamán 1976:93]

with negation (*-chu* marks focus, predicative element):

(86) *Mana -m huwis -chu ñuqa -qa ka -ni.*
Not -DIRE judge -NEG I -TOP be -1.SG
'I am not a judge (my profession is something else).'[a]

(87) *Mana -m ñuqa -chu huwis -qa ka -ni.*
Not -DIRE I -NEG judge -TOP be -1.SG
'The judge is not me (the judge is someone else).'

[Cusihuamán 1976:93]

[a] *huwis*: from Spanish *juez* - 'judge'

Since the distribution of the discourse-relevant morphology is highly regular in equational clauses, the most efficient approach to insert topic and focus markers into this type of clause is a set of rules.

For other clause types however, correctly placing topic and focus markers is a far more difficult task, that will be elaborated in more detail in the following section.

Focalized		Topicalized	
pred	302	subj	259
subj	60	mod	35
linker	12	tmp	23
co	11	linker	19
mod	11	loc	9
caus	8	ben	7
tmp	7	purp	6
sub	7	par	4
comp	2	sub	3
purp	1	acmp	1
loc	1	instr	1
instr	1	voc	1

TABLE 5.9: Distribution of Topic and Focus in Equational Clauses in the SQUOIA Treebank

5.8.2 Modeling Information Structure for Machine Translation

In order to model the information structure in the target language Quechua in cases that cannot be handled by rules (see previous section 5.8.1), we have two basic options:

1. determine topic and focus in the source language Spanish and include the relevant Quechua suffixes during transfer
2. look only at the Quechua output and predict the focalized and topicalized elements

The problem with the first option is that Spanish, unlike Quechua, marks information structure at least in part through intonation and stress. However, since the translation system works only with written text, and we do not have access to Spanish data annotated with the relevant information, we cannot extract topic and focus from the source text. For a small number of fixed expressions, such as *en cuanto a* - 'as for', that clearly introduce a topicalized element, we can rely on rules. Likewise, we can insert topic and focus in *wh*-questions and in equational clauses, if we assume the most frequent distribution of subject as topic and predicative element as focus as always correct.

The second option, to learn and predict the distribution of the discouse markers on Quechua data, seems more promising, especially since topic and focus are annotated in the SQUOIA treebank. Table 5.10 illustrates the distribution of topic and focus in relation to the dependency labels. As becomes evident, the subject is by far the most

Topicalized		Focalized	
subj	766	pred	311
mod	160	subj	198
linker	145	mod	170
tmp	109	sntc	151
obj	83	linker	145
loc	62	obj	137
sub	53	sub	81

TABLE 5.10: Most Frequent Dependency Labels on Topic and Focus in the SQUOIA Treebank

frequently topicalized element, but it is also a very likely candidate for focus. Since finding topic and focus in a general approach that considers the whole sentence is hardly feasible for the reasons mentioned above, we can reduce the problem for the translation system to the decision if the subject is marked for topic, focus or none. This approach is not a complete solution to place the discourse-relevant morphology in the output of the translation system, since it will only cover the most frequent cases.

The setup for the classification is similar to the approach for the disambiguation of subordinated verb forms (see section 5.3.3.1): we use libsvm to train a classifier that predicts for a given subject if it should be marked for topic or focus, or not be marked at all. The SQUOIA treebank contains 1979 sentences with a total of 1427 overt subjects in finite clauses. In order to avoid using lemmas for the classification, we take the Spanish translation of the Quechua roots and use the corresponding semantic features from the Spanish wordnet [Gonzalez-Agirre et al. 2012] and the semantic noun lexicon of the Spanish Resource Grammar [Marimon et al. 2007], the same resources that we used for the verb disambiguation. However, this is an imprecise approach: the treebank contains only the Spanish translation of the Quechua root, but not the lemma, and the Spanish translation can change drastically with morphology.[32] The subject might even be a nominalized verb – in this case, the translation of the root is a Spanish verb, not a noun. Furthermore, the morphological annotation only contains the most frequent translation of the Quechua root, which might not be the best translation in the actual

[32] e.g. *wasi* -'house', is an inanimate object or a location, whereas *wasiyuq* - 'the house-owner' is a person. Regardless of this change in semantics, the treebank only contains the translation of the root, which is 'house'.

context. A better approach would be possible if our treebank was aligned on word level to its Spanish counterpart, since alignments would indicate the correct Spanish lemma in the given context, but unfortunately we do not have these annotations.[33] Furthermore, we tested if semantic information about the head verb of the clause helps to decide on the topic status of the subject.

In addition to the semantic features of the subject, we include a number of more general attributes:

1. does the root occur in the preceding sentence?
2. does the root occur in the following sentence?
3. does the root occur more than once in this sentence?
4. is the word capitalized?
5. does the subject have a dependent that is a demonstrative pronoun?
6. is the subject a pronoun?
7. (does the subject precede the verb?)

All these features were extracted for the subjects of finite verbs in the treebank, which results in 1427 instances for training. The first idea was to let the classifier assign 3 possible values: topic, focus or none. However, the treebank contains only 198 subjects marked for focus, which is not enough for the classifier to learn the distinction reliably. For this reason, we reduced the task for the classifier to the binary decision if a subject is topic or not.

Table 5.11 contains the results of the classification with different features. Note that the information as to whether the subject precedes the verb (feature 7) improves the classification by 1-2%. Unfortunately, we cannot rely on this information during the translation process: since the reordering is based on rules, word order follows a strict template that always assumes basic word order. Therefore, in the output of the translation system the subject will always precede the verb and we cannot use this feature for the classification.

[33] An even better solution would be to use a semantic description of the Quechua words instead of relying on the Spanish translation, but there are no semantic lexica available for Quechua.

The size of the training set with 1427 instances is very small, therefore we tested if we could improve the results by adding data from automatically parsed texts. In order to minimize parsing errors, we parsed the training set used for the morphological disambiguation from section 2.4, since this data has been manually disambiguated on morphological level.[34] Since cross-validation on noisy data is not conclusive, we extracted 150 randomly chosen instances from the treebank data as a test set. In order to assess the performance of the classifier trained on the remaining treebank and the additional data from the automatically parsed texts, we trained a second classifier on the treebank data alone, minus the instances used in the test set. As Table 5.12 clearly shows, more, but noisy data does not improve the classification: the scores in cross-validation increase, but since part of the data is probably incorrect, these numbers are not reliable. The relevant results on the test set show clearly that the classifier trained on pure treebank data outperfoms the classifier enhanced with parsed data.

The results are not good enough to insert the topic marker reliably during the machine translation. Apart from the small training set, the semantic features of the subject are imprecise or might even be wrong in some cases since we take the Spanish translation from the morphological annotation of the treebank instead of describing the actual Quechua subject semantically. Furthermore, we lack relevant information about word order, because the translation is always generated in basic word order.

Since assigning topic status to subject does not work reliably enough for a good translation, it is included only as an optional feature: by default, assigning topic status to subjects through libsvm is turned off, but it can be activated with a command line argument.

In summary, our translation system will introduce topic and focus markers through rules in cases where discourse status can be easily inferred from the Spanish text (*en cuanto a, en lo que concierna..*), and also in certain clause types, such as equational clauses and questions with *wh*-words, where the distribution is more or less fixed. In other contexts, the libsvm classifier can optionally assign topic status to the subject, albeit with a high rate of uncertainty.

[34] Two chapters of the autobiography of Gregorio Condori were deleted from this data set, since they are also part of the treebank and thus already contained in the original training set.

discourse status of subjects with libsvm (standard settings): training data: only treebank (1427 instances)		
	3-way distinction: topic-focus-none	
Features	10-fold cv	word order included (feature 7)
SRG noun classes nominal wordnet class wordnet class of main verb semantic roles of subject for main verb synt. features (without order)	56.97	58.37
SRG noun classes nominal wordnet class synt. features (1-6)	56.97	58.86
SRG noun classes synt. features (1-6)	**57.11**	58.79
synt. features (1-6)	51.50	51.93
	binary distinction: topic-none	
Features	10-fold cv	word order included (feature 7)
SRG noun classes nominal wordnet class wordnet class of main verb semantic roles of subject for main verb synt. features (1-6)	60.67	61.18
SRG noun classes nominal wordnet class synt. features (1-6)	61.18	61.32
SRG noun classes synt. features (1-6)	**61.25**	61.95
synt. features (1-6)	59.00	59.00

Table 5.11: Subject Classification with LibSVM

discourse status of subjects with libsvm (standard settings): binary distinction: topic-none			
Features	trained on A,B 10-fold cv	trained on A,B test set C	trained on B test set C
SRG noun classes nominal wordnet class wordnet class of main verb semantic roles of subject for main verb synt. features (without order)	71.09	55.63	**57.62**
SRG noun classes nominal wordnet class synt. features (1-6)	71.09	55.63	**59.6**
SRG noun classes synt. features (1-6)	70.55	54.97	**58.94**
synt. features (1-6)	68.91	55.63	**56.95**

A: training data: train set of morphological disambiguation (see section 2.4)
automatically parsed (1579 instances)
B: treebank data minus test set C (1277 instances)
C: test set, 150 instances from treebank

TABLE 5.12: Subject Classification with Data from Parsing

5.9 Evaluation of the Machine Translation Output

Assessing the quality of a machine translation system is a complex task, since there is usually not a single correct output that we can match, but instead several valid translation options. Furthermore, there are additional factors to consider apart from grammatical correctness, such as fluency and adequacy.[35] As for methods, there are two fundamentally different approaches: an evaluation can be done by human translators, where each system output is rated according to a predefined scale, or it can be done automatically through a comparison with one or more human translations of the same text. The problem with the former is that human judges differ in their assessment considerably, and we ideally need a large group of judges to obtain meaningful scores. Manual evaluation is complex and expensive, and for this reason, automatic evaluation metrics are more commonly used for MT evaluation. The most popular among these metrics is BLEU (Bilingual Evaluation Understudy): it calculates similarity of the system output to one or more given reference translations by calculating n-gram matches (typically up to 4-grams). Furthermore, a brevity penalty punishes the translations where words have been dropped. The final BLEU score ranges from 0 to 1, typically indicated as percentage, where 1 (or 100%) would be a perfect translation that correponds exactly to the reference translation.

While BLEU has been shown to correlate with human judgement [Papineni et al. 2002], there are some well known issues: BLEU completely ignores the relative importance of words, but some words are more crucial to the meaning of a sentence than others, e.g. omitting the word *not* in an English translation should be punished harder than a missing *the* [Koehn 2010:229]. As for the evaluation of the SQUOIA system with BLEU, we face two main problems: first of all, human Spanish-Quechua translations are usually free translations, they convey roughly the same meaning but cannot be considered close equivalences. Furthermore, information is often distributed differently across sentences. If the original text was written in Quechua and translated to Spanish, pieces of information might be missing in the Spanish text, and thus the automatic translation system, working only on the Spanish text, cannot reproduce the original content of the Quechua

[35] As defined in Koehn [2010:218], fluency rates grammatical correctness and idiomatic word choices, while adequacy assesses if the output of the system conveys the same meaning as the original text, i.e. whether content has been lost, added or distorted.

text. This becomes especially evident in the complex use of the many directional and inter-personal suffixes: for instance, the verbal suffix *-yku* can intensify an action, imply affection, or indicate an inwards movement. Unless the lexical translation of the Spanish verb includes the suffix *-yku*, we cannot insert this piece of information during the translation process. Another obvious example is evidentiality: since the Spanish translation does not contain information about the source of knowledge, the translation system cannot distinguish between direct and indirect evidentiality.

A more technical problem with BLEU comes from the treatment of word forms as basic units to calculate n-gram precision: for agglutinative languages this is bad, since the lack of one suffix in a word form can drastically change the evaluation. Consider the translation example in Fig. 5.14 from page 119:

(88) *Tengo que lavar mi ropa.*
have.1.SG COMP wash my clothes
'I must wash my clothes.'

reference translation: *P'achaymi t'aqsakunay kachkan.*

P'acha -y -mi t'aqsa -ku -na -y ka -chka -n.
clothes -1.SG.POSS -DIRE wash -RFLX -OBL -1.SG.POSS be -PROG -3.SG

'I must wash my clothes.'

[Cusihuamán 1976:210]

MT system output: *P'achay t'aqsanay kan.*

P'acha -y t'aqsa -na -y ka -n.
clothes -1.SG.POSS wash -OBL -1.SG.POSS be -3.SG

'I must wash my clothes.'

None of the words in the translation system output matches the word forms in the reference translation, *P'achay* lacks the evidential marker *-mi*, the verb form *t'aqsanay* lacks the reflexive *-ku* that serves as an intensifier in this case, and the finite verb *kan* lacks the progressive suffix *-chka*. Even though the system output is an acceptable utterance, BLEU calculated on word forms is 0, since there are no n-gram matches. A solution to this problem is to calculate BLEU on n-grams of morphemes instead of word

forms. However, BLEU ideally compares system output to more than one reference translation, but human translated Spanish-Quechua texts are rare, and Spanish texts with several Quechua counterparts are practically non-existent.

Another common evaluation metric for machine translation is TER, the translation edit rate, that counts how many basic edits (insertion, deletion, substitution and shifts) are necessary to convert system output into one of the reference translations.[36] Only the closest reference translation in terms of necessary edit operations is considered for the calculation of the edit rate. While TER is usually based on pre-existing reference translations, a modification thereof, the human-targeted translation edit rate (HTER), is based on corrected system output provided by human judges [Snover et al. 2006].

Therefore, instead of just calculating BLEU with the original Quechua text as reference, we rely on bilingual speakers that manually judge the translation of a short text and optionally provide a corrected Quechua version of the sentence. We can then use the provided corrected translations to calculate BLEU and HTER scores.

5.9.1 Setting

A total of nine bilingual native speakers of Cuzco Quechua or closely related dialects[37] participated in the evaluation of the SQUOIA system. Each of them was presented with the automatic translation of 46 sentences of the story *Catalina y la unkuña mágica - Catalinacha layqay unkuñantin* [Romero Ricalde 2008] and the Spanish input, but not the corresponding Quechua text from the book. The judges were asked to assess the quality of the translation from *bien* - 'good', *se entiende* - 'comprehensible', *mal* - 'bad' to *incomprensible* - 'incomprehensible'. Furthermore, they were asked to provide an alternative translation in case the output of the system was incorrect, see Fig. 5.15 with one of the sentences from the evaluation.

[36] Edit operations are usually calculated on words, but for the evaluation of the Quechua output, we calculated 2 scores, one based on words, and one based on morphemes. The latter should give us a more accurate assessment of the quality, since it will take into consideration missing or superfluous suffixes.

[37] Five speakers of Cuzco Quechua, three speakers of the closely related Abancay dialect and one speaker of Puno Quechua.

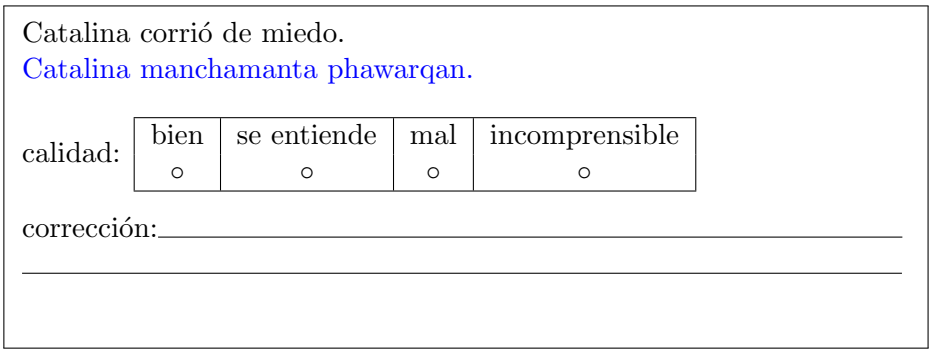


FIGURE 5.15: Evaluation Questionnaire Excerpt

5.9.2 Results

Besides the rating score the judges assigned to the translated sentences, we used the translation corrections from the questionnaires to calculate BLEU and HTER scores. Table 5.13 contains the scores across all evaluators: overall, the system received a rating of 2.5, which is somewhere between 'comprehensible' and 'bad'. However, it turned out that individual sentences received completely different scores by the judges: for instance, the translation of the sentence *Catalina corrió de miedo - Catalina manchamanta phawarqan* from Fig. 5.15 was scored by one judge as 'good', but by another judge as 'incomprehensible'. Hence, the rating scores alone cannot serve as an absolute value for the quality of the translations, but they allow us to compare the translation quality of the individual sentences relative to each other. Both BLEU and HTER indicate how close a system translation is compared to the correction provided by the judge. However, there is an important difference to keep in mind: BLEU measures the similarity of n-grams, hence the higher the BLEU score, the better the translation should be. HTER on the other hand measures the number of necessary edits, as a consequence, the lower the HTER score, the better the translation hypothesis. As expected, BLEU scores calculated on morphemes (63.13) are higher than the numbers obtained through n-gram matching of whole word forms (57.98).

Figure 5.16 illustrates the rating scores and HTER for the individual judges, [38] and Figure 5.17 contains the BLEU scores calculated on the corrections. As becomes evident, the ranking scores do not always correlate with BLEU and HTER: even though judge 3 was relatively lenient in scoring the translations (overall 2.6), the BLEU score calculated

[38] Judge 4 did not provide rating scores, therefore his value is indicated as 0 in Figure 5.16.

all evaluators	words	morphemes
overall rating (1:good-4:incomprehensible)	2.5	–
HTER	29.5	30.3
BLEU	57.98	63.13

TABLE 5.13: Total Rating, HTER and BLEU Evaluation

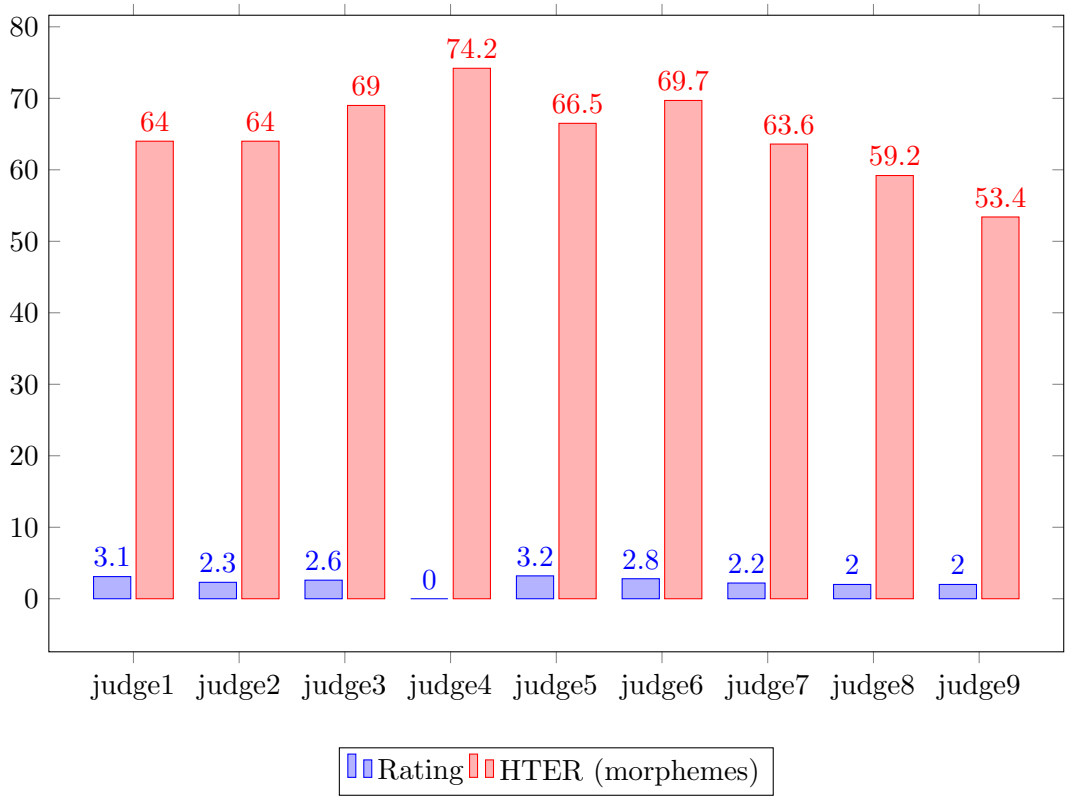


FIGURE 5.16: Rating and HTER Scores of Individual Judges

on his corrections is the lowest of all judges (13.4), and his edit rate score is among the highest (69), which suggests that the corrections he made to the system output were more extensive than those by other judges.

Furthermore, it is important to note that the HTER scores for the individual judges are worse than the overall HTER score of 45.6 calculated on all reference translations: HTER only takes the closest reference translation for each sentence into account, so the HTER score from Table 5.13 is the edit rate of the system output if we compare the system output of each sentence to its closest correction. The HTER values for the individual judges, on the other hand, measure only the edit rate of the system output to their own corrections, which might not be close at all and thus were not considered

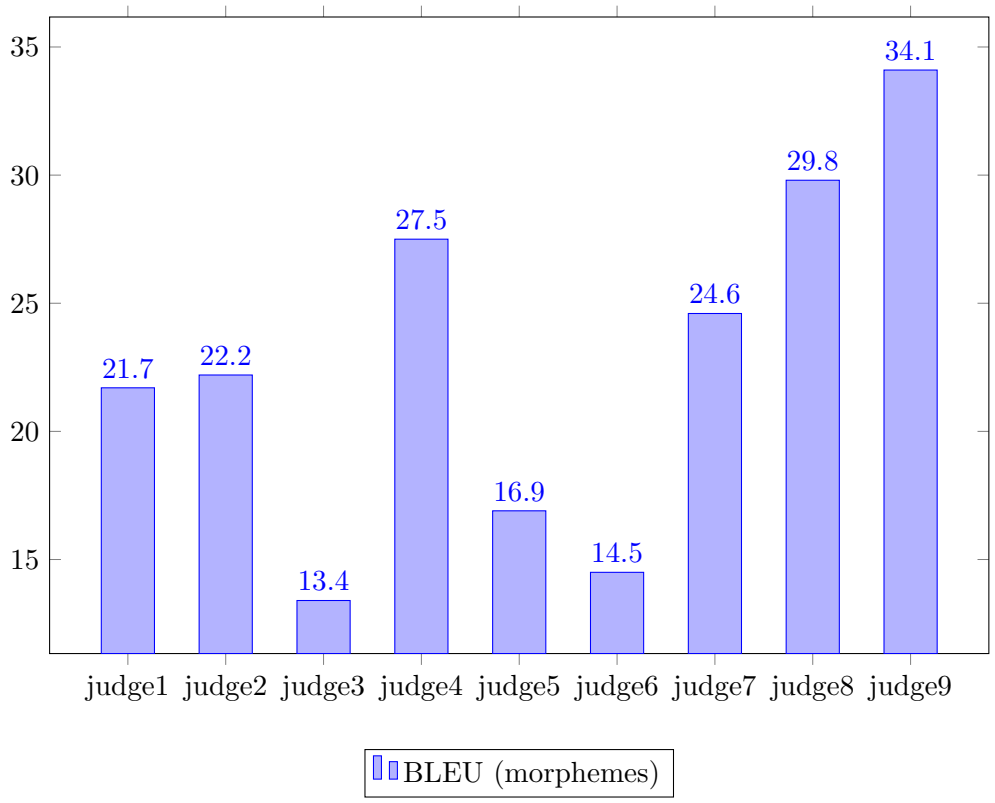


FIGURE 5.17: BLEU Scores of on Corrections provided by Judges

for the overall score in Table 5.13.

Likewise, the BLEU score in Table 5.13 is considerably higher than the BLEU scores calculated on the individual corrections in Table 5.17, since there is a better chance of n-gram matches with nine reference translations as opposed to only one: generally, the more reference translations per sentence, the higher the BLEU score [Papineni et al. 2002].

Table 5.14 lists the most frequent type of errors that occurred in the test set: the most common source for wrong translations are parser errors. These range from relatively small mistakes, such as labeling the subject as object and vice versa, to parse trees that are a complete mess, e.g. a relative clause attached to a wrong head. In the latter case, the translation output was mostly rated incomprehensible, but note that also small parsing errors can lead to ungrammatical sentences that are hard to understand. Another frequent issue concerns errors in the lexical translation of ambiguous words or missing entries in the bilingual dictionary. For instance, the text is about an *unkuña* (a special type of cloth), several times referred to as *tejido* in the Spanish version. The

error types	
wrong parse tree	18
lexical choice	12
wrong case	3
tagging	3

TABLE 5.14: Most Common Error Types

Spanish word *tejido*, like its English counterpart 'tissue', can refer to both textile and organic tissue. In Quechua, on the other hand, there is no single term for both meanings, *tejido* can be translated either as *away* -'cloth' or *aycha* - 'flesh, organic tissue'. The system produces both options, but the correct hypothesis with *away* was scored as slightly less probable by the language model, and since we only used the highest scoring translation for this evaluation, *tejido* was translated as *aycha* in several clauses, resulting in non-sense output. Additionally, wrong case assignment by rules and tagging errors account for a small number of the bad translations. While translations with case errors are wrong, they are usually still comprehensible. Tagging errors, on the other hand, inevitably lead to messy parse trees and usually result in gibberish translations.

Further confirmation as to why parser accuracy is so crucial becomes obvious if we compare ratings by length of the sentence: Figure 5.18 illustrates the negative correlation of sentence length and rating score: the longer a sentence, the worse its score. While there were several short sentences rated as 1=good or 2=comprehensible by most judges, the rating falls drastically with increasing sentence length. The main reason for this negative correlation is that the Spanish parser works relatively reliably on short, simple sentences, but the risk of parsing errors increases with complex, nested sentences. Keep in mind that even though our parser reaches state-of-the-art performance, which is about 90% attachment score, this still means that one out of ten tokens is attached to the wrong head. This is a severe limitation on the quality of the translation output. Furthermore, an obvious drawback of a rule-based core system is that it produces translations that are too literal: unless we have a set of rules that transform a certain structure in the source language into the corresponding structure of the target language, we cannot produce non-literal translations at all.

[39]Sentence length: number of tokens in the Spanish input sentence.

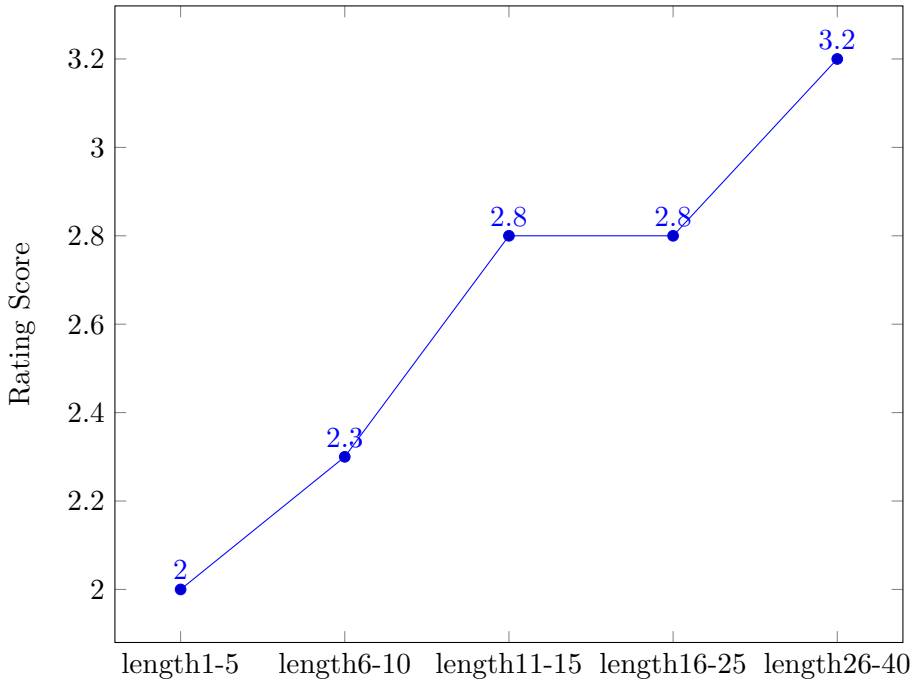


FIGURE 5.18: Average Rating Scores on Sentence Length (Number of Tokens)[39]

5.10 Summary

In this chapter we presented the individual steps in the implementation of our machine translation system for Spanish-Quechua. We illustrated in detail how we improved the automatic analysis of the Spanish input through additional statistics (tagging) and through rules (parsing). Furthermore, we described the problem of generating the correct verb form in subordinated and relative clauses and how we combine machine learning and rules to solve this task. We explained how we built the bilingual dictionary and what kind of information required manual annotation in order to overcome ambiguities resulting from different lexicalization patterns in the source and target language. Lexical disambiguation is done in two steps: a first, rough disambiguation of clear cases by a set of rules immediately follows the lexical transfer, while the more difficult cases are handled by a statistical language model that weighs the resulting translation options at the end of the process. Morphological ambiguities, on the other hand, are resolved by a purely rule-based approach (apart from verb forms). Likewise, syntactic transfer and reordering are covered by an entirely rule-based approach, even though this is not ideal for the latter: since rules handle the reordering, the translations will always have default word order, so changes in word order that might be relevant for discourse structure

are lost in the translation. Furthermore, we experimented with different approaches to insert the discourse-relevant morphemes that mark topic and focus in Quechua. Unfortunately the results are not good enough to include information structure reliably in the translation. A special challenge is set by the evidential suffixes: their placement depends on information structure and clause type, and moreover, their evidential value can in most cases not be inferred from the source sentence, since evidentiality is not a grammatical category in Spanish. The translation system can therefore only provide a basic translation, and a human translator needs to fill in the missing morphemes to make a complete, natural sentence. Our machine translation system has a good lexical coverage and large number of rules to deal with the most common phenomenona. Additionally, the system can handle some of the complex transformations where the target language Quechua differs substantially from the source language Spanish.

Chapter 6

Conclusions

6.1 Recapitulation and Contributions

All applications and resources have been developed over the past 4 years within the SQUOIA project at the University of Zurich, funded by the Swiss National Science Foundation under grants 100015_132219 and 100015_149841. The main part of our work is the translation system for the language pair Spanish-Cuzco Quechua described in Chapter 5. Since the target language Quechua is not only a non-mainstream language in the field of natural language processing, but also typologically quite different from the source language Spanish, several problems became evident and we had to find solutions for them. In some cases, such as the annotation scheme of the dependency treebank, existing resources for typologically similar languages (in this case: Turkish) provided some guidance. For other parts, however, we had to explore completely new paths that sometimes led to success (e.g. the translation of verb forms in subordinated clauses from chapter 5.3.3), while in other cases, the results were not good enough to reliably improve the translation quality (e.g. the insertion of discourse-relevant morphology from chapter 5.8).

Even though the main part of this thesis is about machine translation, Chapter 2 is entirely dedicated to the treatment of Quechua morphology. Morphological analysis and normalization are not directly related to machine translation, but automatic processing of Quechua morphology is needed to create important resources for parts of the translation system. Chapter 2 contains thus an overview of Quechua morphology and

word formation and the systems we developed to automatically process and normalize Quechua texts. Furthermore, we explain briefly how we adapted these resources to create a Quechua spell checker back-end and a spell check plugin for LibreOffice/OpenOffice.

Additionally, the SQUOIA project involved the creation of a Quechua treebank, a collection of texts annotated with syntactic structure and morphological information. Chapter 3 contains a detailed description of the dependency annotation and the difficulties which certain Quechua structures present for this kind of syntactic analysis. The treebank provided not only training data for some of the translation modules, but also served as a source of verification, since it makes it possible to assess the distribution of certain syntactic or morphological structures. This application scenario will also be interesting for Quechua linguists.

A side product of the parallel corpus is the Quechua-Spanish version of the concordance tool Bilingwis presented in Chapter 4: the parallel Quechua-Spanish texts are aligned on word- and morpheme-level and can be searched through a web-interface for translations in context. This application does not have any direct use for the automatic translation system, it might however be interesting for human translators or language learners, since it allows users to look up how a certain word has been translated in other contexts.

6.2 Discussion and Research Questions

We started this thesis with a set of research questions that we will now seek to answer. Our initial question, *How much do we have to rely on rules in NLP applications that involve a low-resource language and can we still make use of statistics?*, cannot be answered with an exact quantification, although our work shows that purely rule-based solutions have severe limitations: we can analyze Quechua word forms with a rule-based finite-state system and achieve high coverage, but in order to get the correct analysis of ambiguous word forms in a given context, statistics provide an easier method than rules, since we cannot write rules to cover all contexts. Likewise, for lexical disambiguation in machine translation, rules are unpractical, since the effort to write enough rules to achieve decent coverage is far too high. On the other hand, statistical models that rely on word counts are not suited well for languages with rich morphology, since a single root can occur in thousands of different word forms. Furthermore, the lack of standardization

in written text hampers any statistical approach. Therefore, even though statistics are better suited for certain problems, a considerable amount of manual work was necessary in order to get clean data for training.

Our second research question, *What implications does a complex agglutinative morphology have for the development of NLP applications and to what extent do we need to adapt common approaches?* has thus already been answered in part: statistical models that rely on counts of word forms suffer from sparse data in morphologically rich languages. For many applications, however, we can overcome this problem by splitting the word forms into smaller units: instead of building the treebank on tagged word forms, we use morphological analysis and disambiguation to split the words into morpheme groups, which are the basic units for the syntactic annotation. Similarly, the statistical language model used for the lexical disambiguation was trained on morphemes instead of words, as a means to overcome the data sparseness.

The third research question, *What issues arise in a machine translation system with typologically quite distinct source and target language that encode different grammatical categories?* deserves particular attention: Quechua is not only typologically quite distinct from Spanish, it also has some characteristics that have rarely (or not at all) been dealt with in machine translation. One of the most important issues concerns verb forms in subordinated clauses: while Spanish has mostly finite verbs in subordinated clauses that are marked for tense, aspect, modality and person, Quechua often has nominal forms that vary according to the clause type. Furthermore, Quechua uses switch-reference as a device of clause linkage, while in Spanish, coreference of subjects is unmarked and pronominal subjects are usually omitted ('pro-drop'). This leads to ambiguities when Spanish text is translated into Quechua.

Another special case is that of relative clauses: the form of the nominalized verb in the Quechua relative clause depends on whether the head noun is the semantic agent of the relative clause. In Spanish, on the other hand, relative clauses can be highly ambiguous, as in certain cases relativization on subjects and objects are not formally distinguished, but instead require semantic knowledge to understand. Furthermore, Spanish can express possession in a relative clause with *cuyo* - 'whose', while there is no such option in Quechua. The translation system can currently not handle this case, as it would require a complete restructuring of the sentence. Another difficult case is

that of translations that involve the first person plural: Spanish has only one form, but Quechua distinguishes between an inclusive ('we and you') and an exclusive ('we, but not you') form. Unless the Spanish source explicitly mentions if the 'you' is included or not, we cannot know which form to use in Quechua and thus generate both. The user will have to choose which form is appropriate.[1]

Furthermore, Quechua conveys information structure in discourse not only through word order, but also through morphological markings on *in situ* elements, while in Spanish, information structure is mostly expressed through non-textual features, such as intonation and stress. We have experimented with machine learning to insert discourse-relevant morphology into the Quechua translation, but the results are not good enough to be used reliably for machine translation. Apart from a few cases that allow a rule-based insertion, we cannot include the suffixes that mark topic and focus in the Quechua translation.

However, the most challenging issue regarding different grammatical categories for the translation system is evidentiality (data source marking): while Spanish, like every language, does have the means to express the source of knowledge for an utterance, evidentiality is not a grammatical category in this language, so explicit mention of the data source is optional and usually absent. In Quechua, on the other hand, it is often claimed that evidentiality needs to be expressed for every statement. Unmarked sentences are actually possible in discourse, but they are usually understood as the speaker having direct evidence, unless the context clearly indicates indirect evidentiality [Faller 2002:168]. Since evidentiality encodes a relation of the speaker (or writer) to his proposition, and thus requires knowledge about the speaker and his experience in the world, this information cannot be inferred from the Spanish source text.

As for our last research question, *What are the essential resources to build a machine translation system with a low-resource target language and how can they be created efficiently?*, some of the steps that were necessary for the machine translation into Quechua

[1] On a side note, one of the texts in the treebank that was translated by a native speaker into Quechua, the *Festschrift 40th anniversary of the Peruvian-German chamber of commerce and industry* (see chapter 3.2) contains many first plural forms, since the writers, members of the chamber, commemorate important events of the organization's history. As for the translation, it is hard to decide whether the inclusive or exclusive form is appropriate: is the reader supposed to be the addressee, and is this reader understood as an outsider or as a co-member of the chamber? In the first case, we want the exclusive form in the translation, but in the latter case, the inclusive form is correct. The Quechua translator actually switches between inclusive and exclusive forms, which may indicate that this ambiguity in a Spanish text can be hard to resolve even for a native speaker.

might apply to other languages in similar situations. As mentioned above, any statistical processing that relies on word (or morpheme) forms needs standardized written text. Automatic normalization is thus essential for languages that either lack a written standard or with several competing standards, i.e. if a single word form appears in different spellings across texts. Furthermore, for agglutinative or even polysynthetic languages, it is beneficial to use smaller units than word forms as the basis for statistical processing, in order to avoid data sparseness. For language pairs where only the target language is low in resources, ambiguities that arise during the translation due to different information being encoded in source and target language can often be solved more easily in the source language. In the SQUOIA translation system, this includes the disambiguation of verb forms based on semantic lexica and treebanks available for Spanish. It is important to notice, however, that certain information might simply not be available from standard sources, and thus has to be created manually. For instance, many Spanish verbs that in their unmarked form are transitive are lexicalized as intransitive verbs in Quechua: in Spanish, the intransitive form is marked, while in Quechua, the transitive form is marked. Even though there are numerous Spanish-Quechua dictionaries, none contains an explicit matching of verb argument structures. We had to manually insert this information into the bilingual dictionary, since the translation of many Spanish verbs depends on their transitivity.

As a concluding remark, it must be emphasized that Quechua, even though a clearly low-resourced language from a computational linguist's point of view, is a linguistically well documented language: numerous grammars have been published for different varieties and linguists have described all kinds of phenomena in the Quechua languages. Furthermore, printed texts, although limited in number and diversity, are available, especially in the Southern Quechua dialects that are relevant for this project. As for machine translation into languages that are less well documented and for which not even a small number of written text is available, the implementation of a machine translation system is certainly more difficult or even impossible.

6.3 Outlook

We have developed a basic language technology toolkit for Cuzco Quechua that includes automatic morphological analysis, spell checking, disambiguation and normalization of

texts. Furthermore, we provide a treebank annotated with dependency trees and a model for MaltParser to automatically annotate the syntactic structure of Quechua texts, and scripts to convert between the different formats. The largest part of this work, however, deals with the implementation of a hybrid machine translation system for the language pair Spanish-Quechua. All tools and resources described in this thesis are published open source under Apache License Version 2.0 and can be used and adapted freely by interested parties.[2].

In this section, we address some of the open issues and options for improvement or enhancement of our basic language technology toolkit.

6.3.1 Morphology Tools

The morphological analysis presented in Chapter 2 already has high coverage, but so far, the lexicon entries are mostly from the Ayacucho and Cuzco varieties of Quechua. The analyzers would profit from the inclusion of Bolivian Quechua and Puno Quechua roots. For Quechua dialects outside the Southern group (QIIC), such as the varieties spoken in Central and Northern Peru and in Ecuador, the morphological analysis presented here is not accurate. Nonetheless, the tools presented here could provide a starting point to implement similar applications for those distant Quechua varieties.

6.3.2 Treebank

The number of annotated sentences in the SQUOIA treebank presented in Chapter 3 is still relatively small. However, we have trained MaltParser on our treebank and offer a complete parsing package to automatically analyze and parse Quechua texts.[3]. Additionally, we provide several scripts and tools to convert between the different formats. With our parsing pipeline, future annotations will be less time-consuming and interested parties can build their own treebank automatically, correct the resulting trees, and thus add to the amount of available annotated Quechua text.

[2]Source code and packages are available from GitHub: https://github.com/ariosquoia/squoia

[3]Individual packages for morphological analysis and parsing are available from: https://github.com/ariosquoia/squoia/releases

6.3.3 Bilingwis

The back-end of Bilingwis is currently undergoing major improvements. The new Bilingwis system will make it possible to search for sequences of tokens, as opposed to individual words (or morphemes in Quechua) in the current version. Once this revision is completed, we will add more parallel text to the web application.

6.3.4 Machine Translation

Our machine translation system presented in Chapter 5 can reliably translate short, simple sentences, but for long and complex sentences, parser errors have a severe impact on the translation quality. A first step has already been taken: we have switched to MaltParser in our translation pipeline. Even though this change was necessary for purely technical reasons, the parsing accuracy has improved since we trained the new model on a larger training set.[4]

There are several open issues regarding machine translation from Spanish to Quechua: above all, the insertion of evidentiality into the Quechua output has not been addressed yet, and it is unclear whether an automatic approach is even possible in this case: evidentiality is a grammatical category in Quechua that requires the source of information to be indicated for each statement.[5] In simplified terms, if the speaker or writer has personally experienced what he describes, he uses direct evidentiality. If, on the other hand, he heard the information from someone else, he uses indirect evidentiality. Since the source of knowledge thus depends on the relation of the speaker to his proposition, and Spanish does not encode this information, we cannot automatically insert the evidential markers. The current system per default translates with direct evidentiality, but it has a switch that can be set at run-time to produce text marked for indirect evidentiality.

Another open issue is the insertion of discourse-relevant morphology into the Quechua translation: the approach presented in this thesis does not work reliably enough to improve the translation. This might be improved with more training data, i.e. more

[4] We used the complete AnCora treebank for training, which amounts to $\sim$ 17,000 sentences. Small adaptions to the original tokenization in the AnCora dependency treebank were necessary in order to ensure consistency with our tagging pipeline with Wapiti and FreeLing as described in section 5.2.

[5] The only exception are imperative clauses.

manually annotated (or at least corrected) dependency trees. Furthermore, alignment on the word level of the Spanish and Quechua parallel treebanks would provide more accurate features for the classifier, as opposed to the Spanish translations of the root in the Quechua trees that were used in the approach presented here.

Finally, the translaton system can always be improved by enhancing the lexical and grammatical coverage: the bilingual lexicon still has many entries that lack a translation and will be translated with the Spanish root, therefore, filling the gaps in the dictionary is an important part of improving the translation system. However, this task is not trivial, since the entries have to follow the format of the lexicon and need to be annotated with morphological and syntactic information. As for the grammatical coverage, the grammar covers the most frequent phenomena, but there are still certain structures that the current system cannot translate, such as relative clauses with *cuyo* - 'whose'.

In summary, our hybrid machine translation is a basic translation engine, it produces Quechua text that needs to be post-edited by a human translator in order to get a fully natural, correct translation. However, we hope that the availability of a machine translation system for Quechua will help translators by creating a raw translation and thus speeding up their work flow.

Appendix A

Machine Translation XML Output

Si no me das el libro que compré en Lima, ya no hablaré contigo.

```
<corpus>
  <SENTENCE ord="1">
    <CHUNK type="grup-verb" si="top" ord="14">
      <NODE ord="14" form="hablaré" lem="hablar" pos="vm" cpos="v"
     rel="sentence" mi="VMIF1S0"/>
      <CHUNK type="grup-verb" si="ao" ord="4">
        <NODE ord="4" form="das" lem="dar" pos="vm" cpos="v" rel="ao"
     mi="VMIP2S0">
          <NODE ord="1" form="Si" lem="si" pos="cs" cpos="c" head="4"
     rel="conj" mi="CS"/>
        </NODE>
        <CHUNK type="sadv" si="mod" ord="2">
          <NODE ord="2" form="no" lem="no" pos="rn" cpos="r" rel="mod"
     mi="RN"/>
        </CHUNK>
        <CHUNK type="sn" si="ci" ord="3">
          <NODE ord="3" form="me" lem="me" pos="pp" cpos="p" rel="ci"
     mi="PP1CS000"/>
        </CHUNK>
        <CHUNK type="sn" si="cd" ord="6">
          <NODE ord="6" form="libro" lem="libro" pos="nc" cpos="n" rel="cd"
     mi="NCMS000">
            <NODE ord="5" form="el" lem="el" pos="da" cpos="d" head="6"
     rel="spec" mi="DA0MS0"/>
          </NODE>
          <CHUNK type="grup-verb" si="S" ord="8">
            <NODE ord="8" form="compré" lem="comprar" pos="vm" cpos="v"
     rel="S" mi="VMIS1S0">
              <NODE ord="7" form="que" lem="que" pos="pr" cpos="p" head="8"
     rel="cd" mi="PR0CN000"/>
            </NODE>
            <CHUNK type="grup-sp" si="cc" ord="9">
              <NODE ord="9" form="en" lem="en" pos="sp" cpos="s" rel="cc"
     mi="SPS00"/>
              <CHUNK type="sn" si="sn" ord="10">
                <NODE ord="10" form="Lima" lem="lima" pos="np" cpos="n"
     rel="sn" mi="NP00G00"/>
              </CHUNK>
            </CHUNK>
          </CHUNK>
        </CHUNK>
        <CHUNK type="F-term" si="term" ord="11">
          <NODE ord="11" form="," lem="," pos="fc" cpos="F" mi="Fc"/>
        </CHUNK>
      </CHUNK>
      <CHUNK type="sadv" si="cc" ord="12">
        <NODE ord="12" form="ya" lem="ya" pos="rg" cpos="r" rel="cc" mi="RG"/>
      </CHUNK>
      <CHUNK type="sadv" si="mod" ord="13">
        <NODE ord="13" form="no" lem="no" pos="rn" cpos="r" rel="mod"
     mi="RN"/>
      </CHUNK>
      <CHUNK type="sn" si="creg" ord="15">
        <NODE ord="15" form="contigo" lem="contigo" pos="pp" cpos="p"
     rel="creg" mi="PP2CS000"/>
      </CHUNK>
      <CHUNK type="F-term" si="term" ord="16">
        <NODE ord="16" form="." lem="." pos="fp" cpos="F" mi="Fp"/>
      </CHUNK>
    </CHUNK>
  </SENTENCE>
</corpus>
```

FIGURE A.1: XML Syntax Tree with Spanish Analysis

Si no me das el libro que compré en Lima, ya no hablaré contigo.

```
<corpus evidentiality="direct">
  <SENTENCE ord="1">
    <CHUNK type="grup-verb" si="top" ord="14" verbform="main">
      <NODE ord="14" form="hablaré" lem="hablar" pos="vm" cpos="v"
     rel="sentence" mi="VMIF1S0"/>
      <CHUNK type="grup-verb" si="ao" ord="4" verbform="main"
     conjLast="chayqa">
        <NODE ord="4" form="das" lem="dar" pos="vm" cpos="v" rel="ao"
     mi="VMIP2S0">
          <NODE ord="1" form="Si" lem="si" pos="cs" cpos="c" head="4"
     rel="conj" mi="CS"/>
        </NODE>
        <CHUNK type="sadv" si="mod" ord="2">
          <NODE ord="2" form="no" lem="no" pos="rn" cpos="r" rel="mod"
     mi="RN"/>
        </CHUNK>
        <CHUNK type="sn" si="ci" ord="3">
          <NODE ord="3" form="me" lem="me" pos="pp" cpos="p" rel="ci"
     mi="PP1CS000"/>
        </CHUNK>
        <CHUNK type="sn" si="cd" ord="6">
          <NODE ord="6" form="libro" lem="libro" pos="nc" cpos="n" rel="cd"
     mi="NCMS000">
            <NODE ord="5" form="el" lem="el" pos="da" cpos="d" head="6"
     rel="spec" mi="DA0MS0"/>
          </NODE>
          <CHUNK type="grup-verb" si="S" ord="8" verbform="rel:not.agentive">
            <NODE ord="8" form="compré" lem="comprar" pos="vm" cpos="v"
     rel="S" mi="VMIS1S0" verbform="rel:not.agentive">
              <NODE ord="7" form="que" lem="que" pos="pr" cpos="p" head="8"
     rel="cd" mi="PR0CN000"/>
            </NODE>
            <CHUNK type="grup-sp" si="cc" ord="9">
              <NODE ord="9" form="en" lem="en" pos="sp" cpos="s" rel="cc"
     mi="SPS00"/>
              <CHUNK type="sn" si="sn" ord="10">
                <NODE ord="10" form="Lima" lem="lima" pos="np" cpos="n"
     rel="sn" mi="NP00G00"/>
              </CHUNK>
            </CHUNK>
          </CHUNK>
        </CHUNK>
        <CHUNK type="F-term" si="term" ord="11">
          <NODE ord="11" form="," lem="," pos="fc" cpos="F" mi="Fc"/>
        </CHUNK>
      </CHUNK>
      <CHUNK type="sadv" si="cc" ord="12">
        <NODE ord="12" form="ya" lem="ya" pos="rg" cpos="r" rel="cc" mi="RG"/>
      </CHUNK>
      <CHUNK type="sadv" si="mod" ord="13">
        <NODE ord="13" form="no" lem="no" pos="rn" cpos="r" rel="mod"
     mi="RN"/>
      </CHUNK>
      <CHUNK type="sn" si="creg" ord="15">
        <NODE ord="15" form="contigo" lem="contigo" pos="pp" cpos="p"
     rel="creg" mi="PP2CS000"/>
      </CHUNK>
      <CHUNK type="F-term" si="term" ord="16">
        <NODE ord="16" form="." lem="." pos="fp" cpos="F" mi="Fp"/>
      </CHUNK>
    </CHUNK>
  </SENTENCE>
</corpus>
```

FIGURE A.2: XML after Verb Disambiguation

```
<corpus evidentiality="direct">
  <SENTENCE ref="1">
    <CHUNK ref="14" type="grup-verb" si="top" verbform="main">
      <NODE ref="14" alloc="" slem="hablar" smi="VMIF1S0" sform="hablaré" UpCase="none" lem="rima" mi="obligative"
      verbmi="VRoot+Obl+1.Sg.Poss">
        <SYN lem="rima" mi="obligative" verbmi="VRoot+Obl+1.Sg.Poss"/>
        <SYN lem="rima" mi="future" verbmi="VRoot+1.Sg.Subj.Fut"/>
        <SYN lem="rima" mi="DS" verbmi="VRoot+DS+1.Sg.Poss"/>
        <SYN lem="rima" mi="SS" verbmi="VRoot+SS+1.Sg.Poss"/>
        <SYN lem="rima" mi="agentive" verbmi="VRoot+Ag"/>
      </NODE>
      <CHUNK ref="4" type="grup-verb" si="ao" verbform="main" conjLast="chayqa">
        <NODE ref="4" alloc="" slem="dar" smi="VMIP2S0" sform="das" UpCase="none" lem="qu" mi="obligative"
      verbmi="VRoot+Obl+2.Sg.Poss">
          <SYN lem="qu" mi="obligative" verbmi="VRoot+Obl+2.Sg.Poss"/>
          <SYN lem="qu" mi="perfect" verbmi="VRoot+Perf+2.Sg.Poss"/>
          <SYN lem="qu" mi="present" verbmi="VRoot+2.Sg.Subj"/>
          <SYN lem="qu" mi="DS" verbmi="VRoot+DS+2.Sg.Poss"/>
          <SYN lem="qu" mi="agentive" verbmi="VRoot+Ag"/>
          <SYN lem="qu" mi="SS" verbmi="VRoot+SS"/>
          <NODE ref="1" alloc="" slem="si" smi="CS" sform="Si" UpCase="none"/>
        </NODE>
        <CHUNK ref="2" type="sadv" si="mod">
          <NODE ref="2" alloc="" slem="no" smi="RN" sform="no" UpCase="none" lem="mana" mi="Part"/>
        </CHUNK>
        <CHUNK ref="3" type="sn" si="ci">
          <NODE ref="3" alloc="" slem="me" smi="PP1CS000" sform="me" UpCase="none" lem="ñuqa" mi="++1.Sg.Obj"/>
        </CHUNK>
        <CHUNK ref="6" type="sn" si="cd">
          <NODE ref="6" alloc="" slem="libro" smi="NCMS000" sform="libro" UpCase="none" lem="unspecified" mi="NRoot">
            <NODE ref="5" alloc="" slem="el" smi="DA0MS0" sform="el" UpCase="none"/>
          </NODE>
          <CHUNK ref="8" type="grup-verb" si="S" verbform="rel:not.agentive">
            <NODE ref="8" alloc="" slem="comprar" smi="VMIS1S0" sform="compré" UpCase="none" lem="ranti"
      mi="indirectpast" verbmi="VRoot+IPst+1.Sg.Subj">
              <SYN lem="ranti" mi="indirectpast" verbmi="VRoot+IPst+1.Sg.Subj"/>
              <SYN lem="ranti" mi="directpast" verbmi="VRoot+NPst+1.Sg.Subj"/>
              <SYN lem="ranti" mi="obligative" verbmi="VRoot+Obl+1.Sg.Poss"/>
              <SYN lem="ranti" mi="perfect" verbmi="VRoot+Perf+1.Sg.Poss"/>
              <SYN lem="ranti" mi="DS" verbmi="VRoot+DS+1.Sg.Poss"/>
              <SYN lem="ranti" mi="agentive" verbmi="VRoot+Ag"/>
              <SYN lem="ranti" mi="SS" verbmi="VRoot+SS"/>
              <NODE ref="7" alloc="" slem="que" smi="PR0CN000" sform="que" UpCase="none"/>
            </NODE>
            <CHUNK ref="9" type="grup-sp" si="cc">
              <NODE ref="9" alloc="" slem="en" smi="SPS00" sform="en" UpCase="none" lem="" adpos="en"/>
              <CHUNK ref="10" type="sn" si="sn">
                <NODE ref="10" alloc="" slem="lima" smi="NP00G00" sform="Lima" UpCase="first" unknown="transfer"/>
              </CHUNK>
            </CHUNK>
          </CHUNK>
        </CHUNK>
        <CHUNK ref="11" type="F-term" si="term">
          <NODE ref="11" alloc="" slem="," smi="Fc" sform="," UpCase="none" unknown="transfer"/>
        </CHUNK>
      </CHUNK>
      <CHUNK ref="12" type="sadv" si="cc">
        <NODE ref="12" alloc="" slem="ya" smi="RG" sform="ya" UpCase="none" lem="ña" moveToHead_mi="++Disc"
      func="adverbial"/>
      </CHUNK>
      <CHUNK ref="13" type="sadv" si="mod">
        <NODE ref="13" alloc="" slem="no" smi="RN" sform="no" UpCase="none" lem="mana" mi="Part"/>
      </CHUNK>
      <CHUNK ref="15" type="sn" si="creg">
        <NODE ref="15" alloc="" slem="contigo" smi="PP2CS000" sform="contigo" UpCase="none" lem="qam"
      mi="PrnPers+Instr">
          <SYN lem="qam" mi="PrnPers+Instr"/>
          <SYN lem="qam" mi="++2.Sg.Obj"/>
          <SYN lem="qam" mi="PrnPers"/>
        </NODE>
      </CHUNK>
      <CHUNK ref="16" type="F-term" si="term">
        <NODE ref="16" alloc="" slem="." smi="Fp" sform="." UpCase="none" unknown="transfer"/>
      </CHUNK>
    </CHUNK>
  </SENTENCE>
</corpus>
```

FIGURE A.3: XML after Lexical Transfer

Si no me das el libro que compré en Lima, ya no hablaré contigo.

```
<corpus evidentiality="direct">
  <SENTENCE ref="1">
    <CHUNK ref="14" type="grup-verb" si="top" verbform="main">
      <NODE ref="14" slem="hablar" smi="VMIF1S0" sform="hablaré" UpCase="none"
     sem="[+speech]" lem="rima" mi="future" verbmi="VRoot+1.Sg.Subj.Fut"/>
      <CHUNK ref="4" type="grup-verb" si="ao" verbform="main" conjLast="chayqa">
        <NODE ref="4" slem="dar" smi="VMIP2S0" sform="das" UpCase="none"
     sem="[ditrans]" lem="qu" mi="present" verbmi="VRoot+2.Sg.Subj">
          <NODE ref="1" alloc="" slem="si" smi="CS" sform="Si" UpCase="none"/>
        </NODE>
        <CHUNK ref="2" type="sadv" si="mod">
          <NODE ref="2" alloc="" slem="no" smi="RN" sform="no" UpCase="none"
     lem="mana" mi="Part"/>
        </CHUNK>
        <CHUNK ref="3" type="sn" si="ci">
          <NODE ref="3" alloc="" slem="me" smi="PP1CS000" sform="me"
     UpCase="none" lem="ñuqa" mi="++1.Sg.Obj"/>
        </CHUNK>
        <CHUNK ref="6" type="sn" si="cd">
          <NODE ref="6" alloc="" slem="libro" smi="NCMS000" sform="libro"
     UpCase="none" lem="unspecified" mi="NRoot">
            <NODE ref="5" alloc="" slem="el" smi="DA0MS0" sform="el"
     UpCase="none"/>
          </NODE>
          <CHUNK ref="8" type="grup-verb" si="S" verbform="rel:not.agentive">
            <NODE ref="8" slem="comprar" smi="VMIS1S0" sform="compré"
     UpCase="none" lem="ranti" mi="perfect" verbmi="VRoot+Perf+1.Sg.Poss">
              <NODE ref="7" alloc="" slem="que" smi="PR0CN000" sform="que"
     UpCase="none"/>
            </NODE>
            <CHUNK ref="9" type="grup-sp" si="cc">
              <NODE ref="9" alloc="" slem="en" smi="SPS00" sform="en"
     UpCase="none" lem="" adpos="en" case="+Loc"/>
              <CHUNK ref="10" type="sn" si="sn">
                <NODE ref="10" alloc="" slem="lima" smi="NP00G00" sform="Lima"
     UpCase="first" unknown="transfer"/>
              </CHUNK>
            </CHUNK>
          </CHUNK>
        </CHUNK>
        <CHUNK ref="11" type="F-term" si="term">
          <NODE ref="11" alloc="" slem="," smi="Fc" sform="," UpCase="none"
     unknown="transfer"/>
        </CHUNK>
      </CHUNK>
      <CHUNK ref="12" type="sadv" si="cc">
        <NODE ref="12" alloc="" slem="ya" smi="RG" sform="ya" UpCase="none"
     lem="ña" moveToHead_mi="++Disc" func="adverbial"/>
      </CHUNK>
      <CHUNK ref="13" type="sadv" si="mod">
        <NODE ref="13" alloc="" slem="no" smi="RN" sform="no" UpCase="none"
     lem="mana" mi="Part"/>
      </CHUNK>
      <CHUNK ref="15" type="sn" si="creg">
        <NODE ref="15" slem="contigo" smi="PP2CS000" sform="contigo"
     UpCase="none" lem="qam" mi="PrnPers+Instr"/>
      </CHUNK>
      <CHUNK ref="16" type="F-term" si="term">
        <NODE ref="16" alloc="" slem="." smi="Fp" sform="." UpCase="none"
     unknown="transfer"/>
      </CHUNK>
    </CHUNK>
  </SENTENCE>
</corpus>
```

FIGURE A.4: XML after Morphological and Prepositional Disambiguation

Si no me das el libro que compré en Lima, ya no hablaré contigo.

```
<corpus evidentiality="direct">
  <SENTENCE ref="1">
    <CHUNK ref="14" type="grup-verb" si="top" verbform="main" chunkmi=",+Neg">
      <NODE ref="14" slem="hablar" smi="VMIF1S0" sform="hablaré" UpCase="none"
     sem="[+speech]" lem="rima" mi="future" verbmi="VRoot+1.Sg.Subj.Fut"/>
      <CHUNK ref="4" type="grup-verb" si="ao" verbform="main" conjLast="chayqa"
     chunkmi=",+Neg" addverbmi=",++1.Sg.Obj">
        <NODE ref="4" slem="dar" smi="VMIP2S0" sform="das" UpCase="none"
     sem="[ditrans]" lem="qu" mi="present" verbmi="VRoot+2.Sg.Subj">
          <NODE ref="1" alloc="" slem="si" smi="CS" sform="Si" UpCase="none"/>
        </NODE>
        <CHUNK ref="2" type="sadv" si="mod" deleteMorph="+Neg"
     chunkmi="+Neg,+DirE">
          <NODE ref="2" alloc="" slem="no" smi="RN" sform="no" UpCase="none"
     lem="mana" mi="Part"/>
        </CHUNK>
        <CHUNK ref="3" type="sn" si="ci" addverbmi="++1.Sg.Obj" case="+Dat"
     delete="yes">
          <NODE ref="3" alloc="" slem="me" smi="PP1CS000" sform="me"
     UpCase="none" lem="ñuqa" mi="++1.Sg.Obj"/>
        </CHUNK>
        <CHUNK ref="6" type="sn" si="cd" case="+Acc">
          <NODE ref="6" alloc="" slem="libro" smi="NCMS000" sform="libro"
     UpCase="none" lem="unspecified" mi="NRoot">
            <NODE ref="5" alloc="" slem="el" smi="DA0MS0" sform="el"
     UpCase="none"/>
          </NODE>
          <CHUNK ref="8" type="grup-verb" si="S" verbform="rel:not.agentive">
            <NODE ref="8" slem="comprar" smi="VMIS1S0" sform="compré"
     UpCase="none" lem="ranti" mi="perfect" verbmi="VRoot+Perf+1.Sg.Poss">
              <NODE ref="7" alloc="" slem="que" smi="PR0CN000" sform="que"
     UpCase="none"/>
            </NODE>
            <CHUNK ref="9" type="grup-sp" si="cc" case="+Loc">
              <NODE ref="9" alloc="" slem="en" smi="SPS00" sform="en"
     UpCase="none" lem="" adpos="en" case="+Loc"/>
              <CHUNK ref="10" type="sn" si="sn">
                <NODE ref="10" alloc="" slem="lima" smi="NP00G00" sform="Lima"
     UpCase="first" unknown="transfer"/>
              </CHUNK>
            </CHUNK>
          </CHUNK>
        </CHUNK>
        <CHUNK ref="11" type="F-term" si="term">
          <NODE ref="11" alloc="" slem="," smi="Fc" sform="," UpCase="none"
     unknown="transfer"/>
        </CHUNK>
      </CHUNK>
      <CHUNK ref="12" type="sadv" si="cc" delete=",yes">
        <NODE ref="12" alloc="" slem="ya" smi="RG" sform="ya" UpCase="none"
     lem="ña" moveToHead_mi="++Disc" func="adverbial"/>
      </CHUNK>
      <CHUNK ref="13" type="sadv" si="mod" deleteMorph="+Neg"
     chunkmi="+Neg,+DirE,+Disc">
        <NODE ref="13" alloc="" slem="no" smi="RN" sform="no" UpCase="none"
     lem="mana" mi="Part"/>
      </CHUNK>
      <CHUNK ref="15" type="sn" si="creg">
        <NODE ref="15" slem="contigo" smi="PP2CS000" sform="contigo"
     UpCase="none" lem="qam" mi="PrnPers+Instr"/>
      </CHUNK>
      <CHUNK ref="16" type="F-term" si="term">
        <NODE ref="16" alloc="" slem="." smi="Fp" sform="." UpCase="none"
     unknown="transfer"/>
      </CHUNK>
    </CHUNK>
  </SENTENCE>
</corpus>
```

FIGURE A.5: XML after Intra- and Interchunk Syntactic Transfer

Bibliography

Adelaar, W. F. H. (1987). Aymarismos en el quechua de Puno. *Indiana*, 11:223–231.

Adelaar, W. F. H. and Muysken, P. (2004). *The Languages of the Andes*. Cambridge Language Surveys. Cambridge University Press, Cambridge.

Adesam, Y. (2012). *The Multilingual Forest - Investigating High-quality Parallel Corpus development*. PhD thesis, Stockholm University, Stockholm.

Alegria, I., Aranzabe, M., Ezeiza, N., Ezeiza, A., and Urizar, R. (2002). Using Finite State Technology in Natural Language Processing of Basque. In *CIAA '01 Revised Papers from the 6th International Conference on Implementation and Application of Automata*.

Atalay, N. B., Oflazer, K., and Say, B. (2003). The Annotation Process in the Turkish Treebank. In *Proceedings of the 4th International Workshop on Linguistically Interpreteted Corpora (LINC)*, Budapest, Hungary.

Attardi, G. (2006). Experiments with a Multilanguage Non-Projective Dependency Parser. In *Proceedings of the Tenth Conference on Natural Language Learning*, New York, NY.

Ballesteros, M. and Nivre, J. (2015). MaltOptimizer: Fast and Effective Parser Optimization. *Natural Language Engineering*, FirstView:1–27.

Beesley, K. R. and Karttunen, L. (2003). *Finite State Morphology*. CSLI Studies in Computational Linguistics. CSLI Publications, Stanford.

Bejček, E., Kettnerová, V., and Lopatková, M. (2010). Advanced Searching in the Valency Lexicons Using PML-TQ Search Engine. In Sojka, P., Horák, A., Kopeček, I., and Pala, K., editors, *Text, Speech and Dialogue*, volume 6231 of *Lecture Notes in Computer Science*, pages 51–58. Springer, Berlin, Heidelberg.

Carstensen, K.-U., Ebert, C., Ebert, C., Jekat, S., Klabunde, R., and Langer, H., editors (2010). *Computerlinguistik und Sprachtechnologie: Eine Einführung*. Spektrum, Heidelberg, 3. edition.

Castro Mamani, R. and Rios Gonzales, A. (2014). Allin Qillqay! A Free Online Web Spell Checking Service for Quechua. In Ugaz Burga, J. E., Gonzales Sánchez, S. R., and Torres Guerra, C., editors, *Memoria - VI Congreso Internacional de Computación y Telecomunicaciones (COMTEL) 2014*, pages 23–30, Lima. Fondo Editorial de la Universidad Inca Garcilaso de la Vega.

Cerrón-Palomino, R. (1994). *Quechua sureño, diccionario unificado quechua-castellano, castellano-quechua*. Biblioteca Nacional del Perú, Lima.

Cerrón-Palomino, R. (2003). *Lingüística Quechua*. Centro de Estudios Regionales Andinos Bartolomé de las Casas (CBC), Cuzco, 2. edition.

Chang, C.-C. and Lin, C.-J. (2011). LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27.

Chávez Gonzales, O., Castillo Collado, M., and Quintasi Mamani (translators), M. (2002). *Perú Suyupa Hatun Rimanakuynin, Acuerdo Nacional nisqa* . Empresa Peruana de Servicios Editoriales S.A.Segraf - Editora Perú, Lima.

Chrupala, G., Dinu, G., and van Genabith, J. (2008). Learning Morphology with Morfette. In Calzolari, N., Choukri, K., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., and Tapias, D., editors, *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC 2008)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Cusihuamán, A. G. (1976). *Gramática Quechua: Cuzco-Collao*. Gramáticas referenciales de la lengua quechua. Ministerio de Educación, Lima.

de Marneffe, M.-C. and Manning, C. D. (2008). Stanford Dependencies manual. Technical report.

Dedenbach-Salazar Sáenz, S., von Gleich, U., Hartmann, R., Masson, P., and Soto Ruiz, C. (2002). *Rimaykullayki - Unterrichtsmaterialien zum Quechua Ayacuchano*. Dietrich Reimer Verlag GmbH, Berlin, 4. edition.

Eryiğit, G. (2007). ITU Treebank Annotation Tool. In *In Proceedings of the Linguistic Annotation Workshop at ACL 2007*, Prague, Czech Republic.

Faller, M. (2002). *Semantics and Pragmatics of Evidentials in Cuzco Quechua.* PhD thesis, Stanford University.

Faller, M. (2004). The deictic core of non-experienced past in Cuzco Quechua. *Journal of Semantics*, 21:45–85.

Floyd, S. (2011). Re-discovering the Quechua adjective. *Linguistic Typology*, 15(1):25–63.

Gasser, M. (2006). Machine Translation and the Future of Indigenous Languages. In *I Congreso Internacional de Lenguas y Literaturas Indoamericanas*, Temuco, Chile.

Gasser, M. (2011). Computational morphology and the teaching of indigenous languages. In Molina, S. C. and McDowell, J., editors, *Proceedings of the First Symposium on Teaching Indigenous Languages of Latin America*, pages 52–63, Center for Latin American and Caribbean Studies, Indiana University, Bloomington, USA.

Gonzalez-Agirre, A., Laparra, E., and Rigau, G. (2012). Multilingual Central Repository version 3.0: upgrading a very large lexical knowledge base. In *Proceedings of the Sixth International Global WordNet Conference (GWC'12)*, Matsue, Japan.

Hastings, R. (2004). *The Syntax and Semantics of Relativization and Quantification: The Case of Quechua.* PhD thesis, Cornell University.

Heafield, K. (2011). KenLM: Faster and Smaller Language Model Queries. In *Proceedings of the EMNLP 2011 Sixth Workshop on Statistical Machine Translation*, pages 187–197, Edinburgh, Scotland, United Kingdom.

Heggarty, P. (2005). Enigmas en el origen de las lenguas andinas: aplicando nuevas técnicas a las incógnitas por resolver. *Revista Andina*, 40:9–57.

Hulden, M. (2009a). Fast approximate string matching with finite automata. *Procesamiento del Lenguaje Natural*, (43):57–64.

Hulden, M. (2009b). Foma: a finite-state compiler and library. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 29–32, Athens, Greece. Association for Computational Linguistics.

Itier, C. (2007). *El hijo del oso: la literatura oral quechua de la región del Cuzco*. IFEA : Instituto de Estudios Peruanos : Fondo Editorial de la Pontificia Universidad Católica del Perú, Lima.

Koehn, P. (2010). *Statistical Machine Translation*. Cambridge University Press, Cambridge.

Kshetri, N. and Nikhiles, D. (2009). Global Digital Divide. In Khosrow-Pour, M., editor, *Encyclopedia of Information Science and Technology, Second Edition*, pages 1664–1670. IGI Global, Hershey.

Kübler, S., McDonald, R., and Nivre, J. (2009). *Dependency Parsing*, volume 2 of *Synthesis Lectures on Human Language Technologies*. Morgan & Claypool.

Landerman, P. (1991). *Quechua Dialects and their Classification*. PhD thesis, University of California, Los Angeles.

Larico Uchamaco, G. R., Calderón Vilca, H. D., and Cárdenas Mariño, F. C. (2013). Incubation system machine translation Spanish to Quechua, based on free and open source platform Apertium. *CEPROSIMAD*, 2(1):57–65.

Lavergne, T., Cappé, O., and Yvon, F. (2010). Practical Very Large Scale CRFs. In *Proceedings the 48th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 504–513, Uppsala, Sweden. Association for Computational Linguistics.

Lira, J. (1990). *Cuentos del Alto Urubamba*. Centro de Estudios Regionales Andinos Bartolomé de las Casas (CBC), Cuzco.

Marimon, M., Fisas, B., Bel, N., Arias, B., Vázquez, S., Vivaldi, J., Torner, S., Villegas, M., and Lorente, M. (2012). The IULA Treebank. In Calzolari, N., Choukri, K., Declerck, T., Doğan, M. U., Maegaard, B., Mariani, J., Odijk, J., and Piperidis, S., editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC 2012)*, Istanbul, Turkey. European Language Resources Association (ELRA).

Marimon, M., Seghezzi, N., and Bel, N. (2007). An Open-source Lexicon for Spanish. *Procesamiento del Lenguaje Natural*, 39:131–137.

Mohler, M. and Mihalcea, R. (2008). Babylon Parallel Text Builder: Gathering Parallel Texts for Low-Density Languages. In Calzolari, N., Choukri, K., Maegaard,

B., Mariani, J., Odijk, J., Piperidis, S., and Tapias, D., editors, *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC 2008)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Monson, C., Llitjós, A. F., Aranovich, R., Levin, L., Brown, R., Peterson, E., Carbonell, J., and Lavie, A. (2006). Building NLP Systems for Two Resource-Scarce Indigenous Languages: Mapudungun and Quechua. In Calzolari, N., Choukri, K., Gangemi, A., Maegaard, B., Mariani, J., Odijk, J., and Tapias, D., editors, *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC 2006)*, Genoa, Italy. European Language Resources Association (ELRA).

Nivre, J. (2005). Dependency Grammar and Dependency Parsing. Technical Report MSI 05133, Växjö University, School of Mathematics and Systems Engineering.

Nivre, J., Hall, J., Nilsson, J., Chanev, A., Eryigit, G., Kübler, S., Marinov, S., and Marsi, E. (2007). MaltParser: A language-independent system for data-driven dependency parsing. *Natural Language Engineering*, 13(2):95–135.

Oflazer, K. (1996). Error-tolerant Finite State Recognition with Applications to Morphological Analysis and Spelling Correction. *Computational Linguistics*, 22(1).

Padró, L. and Stanilovsky, E. (2012). FreeLing 3.0: Towards Wider Multilinguality. In Calzolari, N., Choukri, K., Declerck, T., Doğan, M. U., Maegaard, B., Mariani, J., Odijk, J., and Piperidis, S., editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC 2012)*, Istanbul, Turkey. European Language Resources Association (ELRA).

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 311–318, Stroudsburg, PA, USA. Association for Computational Linguistics.

Parker, G. J. (1963). Clasificación genética de los dialectos quechuas. *Revista del Museo Nacional, Lima*, (32):241–252.

Pirinen, T. and Lindén, K. (2010). Finite-State Spell-Checking with Weighted Language and Error Models. Building and Evaluating Spell-Checkers with Wikipedia as Corpus. In Sarasola, K., Tyers, F. M., and Forcada, M. L., editors, *7th SaLTMiL Workshop*

on Creation and Use of Basic Lexical Resources for Less-Resourced Languages, pages 13–18, Valletta, Malta. European Language Resources Association (ELRA).

Quiroz Villarroel, A. (2000). *Gramática Quechua*. Ministerio de Educación, Cultura y Deportes, Fondo de las Naciones Unidas para la Infancia (UNICEF), Bolivia.

Rios, A. (2011a). Applying Finite State Techniques to a Native American Language: Quechua. Master's thesis, University of Zurich.

Rios, A. (2011b). Spell checking an agglutinative language: Quechua. In *5th Language and Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics*, pages 51–55, Poznán, Poland. Fundacja Uniwersytetu im. A. Mickiewicza.

Rios, A. (2014). Esquema de anotaciones sintácticas para el Quechua Sureño. Technical report, Institute of Computational Linguistics, University of Zurich.

Rios, A. and Castro Mamani, R. (2014). Morphological Disambiguation and Text Normalization for Southern Quechua Varieties. In *Proceedings of the First Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects (VarDial)*, pages 39–47, Dublin, Ireland. Association for Computational Linguistics.

Rios, A. and Göhring, A. (2012). A tree is a Baum is an árbol is a sach'a: Creating a trilingual treebank. In Calzolari, N., Choukri, K., Declerck, T., Doğan, M. U., Maegaard, B., Mariani, J., Odijk, J., and Piperidis, S., editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC 2012)*, pages 1874–1879, Istanbul, Turkey. European Language Resources Association (ELRA).

Rios, A. and Göhring, A. (2013). Machine Learning Disambiguation of Quechua Verb Morphology. In *Proceedings of the Second Workshop on Hybrid Approaches to Translation*. Association for Computational Linguistics.

Rios, A. and Göhring, A. (2016). Machine Learning applied to Rule-Based Machine Translation. In Costa-jussà, M., Rapp, R., Lambert, P., Eberle, K., Banchs, R. E., and Babych, B., editors, *Hybrid Approaches to Machine Translation*, Theory and Applications of Natural Language Processing, page to appear. Springer International Publishing.

Romero Ricalde, B. (2008). *Catalina y la Unkuña Mágica - Catalinacha layqay unkuñantin*. Centro de Estudios Regionales Andinos Bartolomé de las Casas (CBC), Cuzco, Peru.

Sánchez, L. (2010). *The Morphology and Syntax of Topic and Focus - Minimalist inquiries in the Quechua periphery*, volume 169 of *Linguistik Aktuell - Linguistics Today*. John Benjamins Publishing Company, Amsterdam/Philadelphia.

Schmid, H. and Laws, F. (2008). Estimation of conditional probabilities with decision trees and an application to fine-grained POS tagging. In *Proceedings of the 22nd International Conference on Computational Linguistics*, volume 1 of *COLING '08*, pages 777–784. Association for Computational Linguistics.

Seeker, W., Farkas, R., Bohnet, B., Schmid, H., and Kuhn, J. (2012). Data-driven Dependency Parsing with Empty Heads. In *Proceedings of COLING 2012*, pages 1081–1090.

Snover, M., Dorr, B., Schwartz, R., Micciulla, L., and Weischedel, R. (2006). A Study of Translation Error Rate with Targeted Human Annotation. In *Proceedings of the Association for Machine Transaltion in the Americas*.

Soto Ruiz, C. (1976). *Gramática Quechua: Ayacucho-Chanca*. Gramáticas referenciales de la lengua quechua. Ministerio de Educación, Lima.

Soto Ruiz, C. (2006). *Quechua - Manual de enseñanza*, volume 4 of *Lengua y Sociedad*. Instituto de Estudios Peruanos (IEP), Lima, Peru, 3. edition.

Štěpánek, J. and Pajas, P. (2010). Querying Diverse Treebanks in a Uniform Way. In Calzolari, N., Choukri, K., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., Rosner, M., and Tapias, D., editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC 2010)*, Valletta, Malta. European Language Resources Association (ELRA).

Taulé, M., Martí, M. A., and Recasens, M. (2008). AnCora: Multilevel Annotated Corpora for Catalan and Spanish. In Nicoletta Calzolari, K. C. B. M. J. M. J. O. S. P. D. T., editor, *Proceedings of the Sixth International Language Resources and Evaluation (LREC 2008)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Torero, A. (1964). Los dialectos quechuas. *Anales Científicos de la Universidad Agraria, Lima*, (IV):446–478.

Valderrama Fernandez, R. and Escalante Gutierrez, C. (1977). *Gregorio Condori Mamani - Autobiografía*. Biblioteca de la Tradición Oral Andina. Centro de Estudios Regionales Andinos Bartolomé de las Casas (CBC), Cuzco.

Varga, D., Németh, L., Halácsy, P., Kornai, A., Trón, V., and Nagy, V. (2005). Parallel corpora for medium density languages. In *Recent Advances in Natural Language Processing (RANLP 2005)* , pages 590–596, Borovets, Bulgaria.

Volk, M., Göhring, A., Lehner, S., Rios, A., Sennrich, R., and Uibo, H. (2011). Word-aligned Parallel Text : A New Resource for Contrastive Language Studies. In *Supporting Digital Humanities, Conference 2011*, Copenhagen, Denmark.

Vondřička, P. (2014). Aligning parallel texts with InterText. In Calzolari, N., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014)*, pages 1875–1879, Reykjavik, Iceland. European Language Resources Association (ELRA).